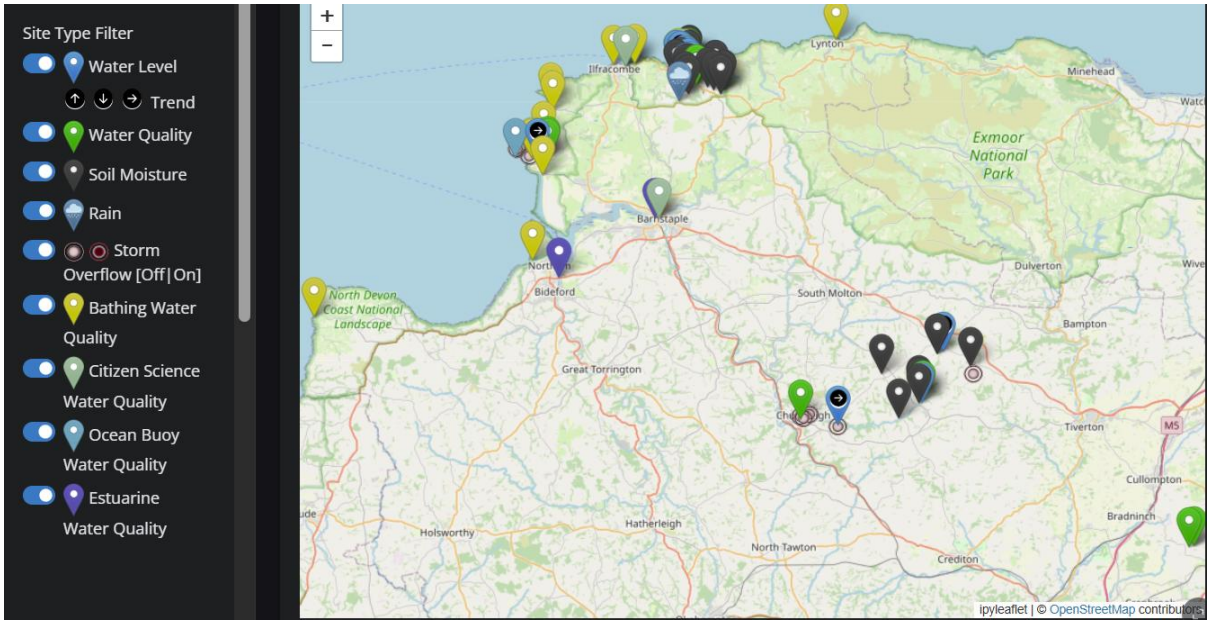
Predicting E. coli risk from continuous sensor data

Last year we began a partnership with the North Devon Biosphere and launched a dashboard exploring continuous sensor data in the region. Since then, several data sources have been added including sample data collected by citizen scientists and Environment Agency (EA) bathing water quality.

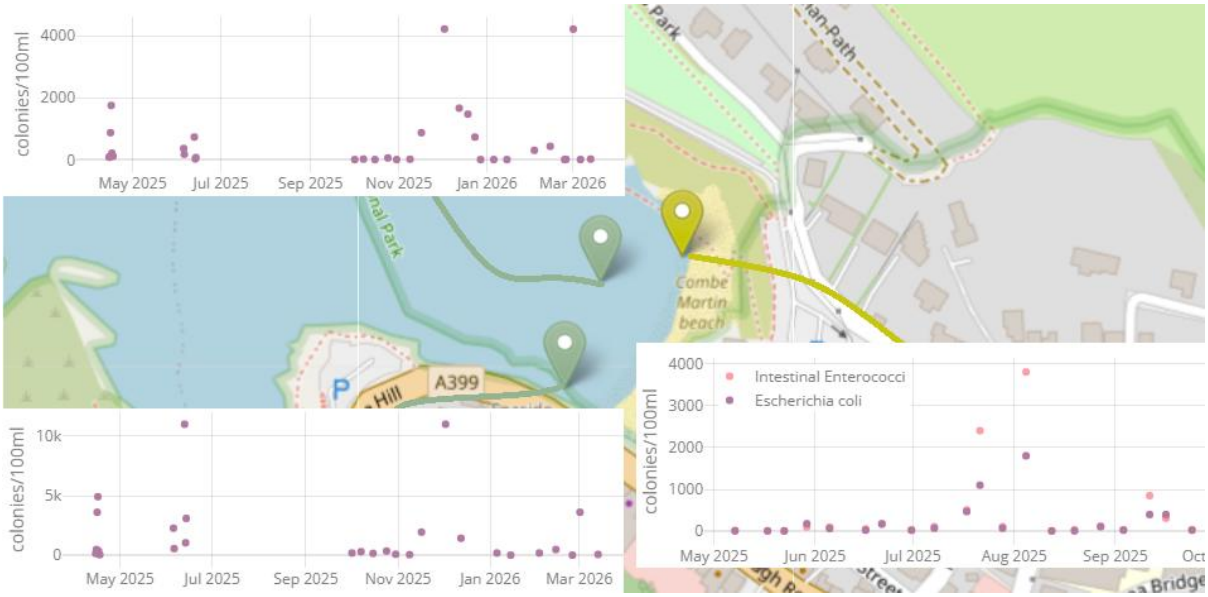


The samples collected are analysed for E. coli (bacterial pollution) as well as other water quality indicators. Volunteers regularly collect water samples at sites and these are analysed using culture-based methods (growing colonies on a plate) - in this case using a portable testing unit rather than sending to a lab. The bacterial load results come back as a count of colony-forming units (cfu) per 100ml. High E. coli levels indicate faecal contamination, which can pose a health risk to people using the water.

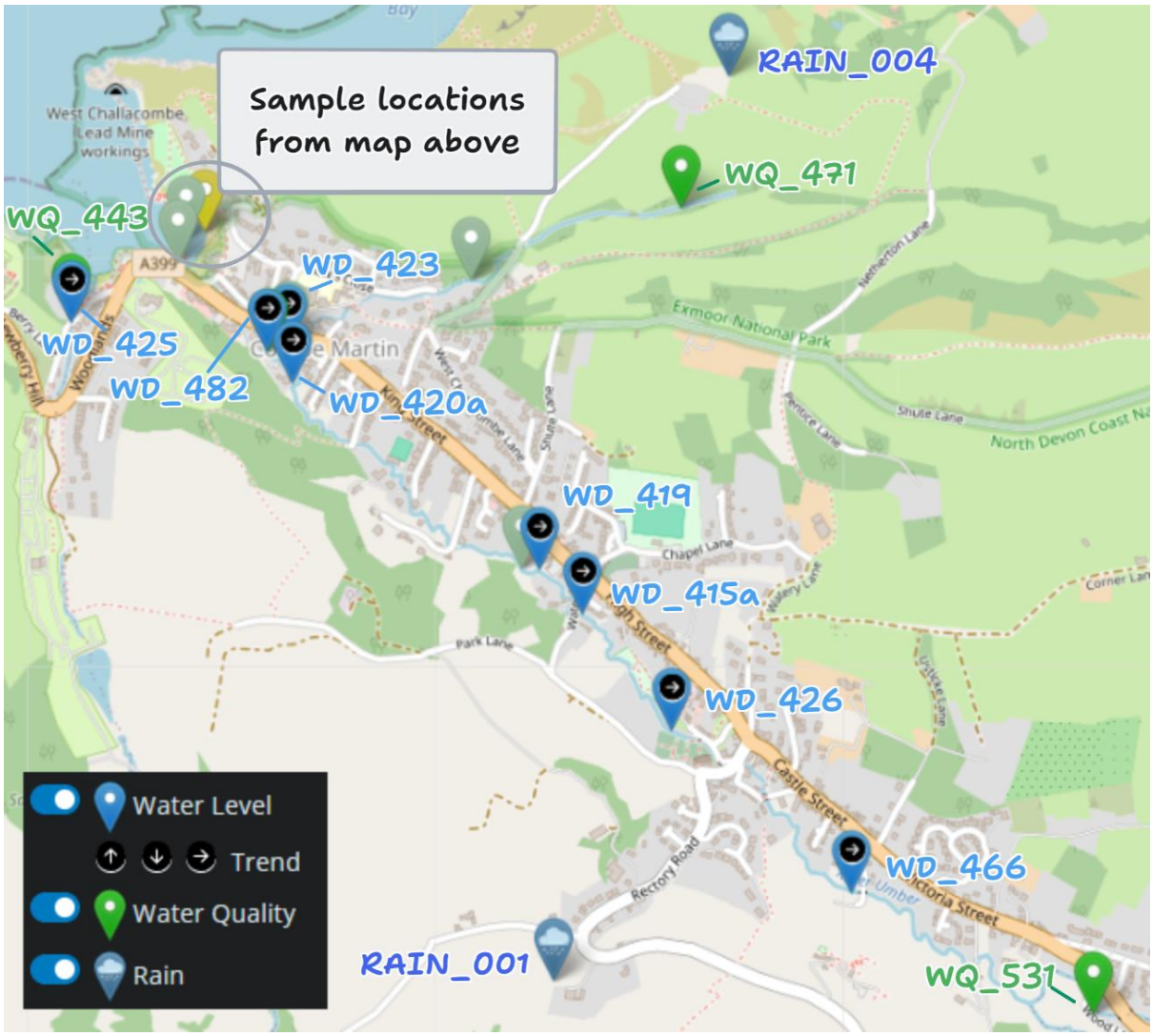
We are interested in whether the E. coli level can be predicted given the continuous sensor data we have for rainfall, water quality and water depth within the catchment. Sampling to measure E. coli is infrequent because it relies on volunteer's time and lab resources or expensive equipment, so there are gaps between measurements and even when you have a sample, there is a lag before you know the result. A model using continuous sensor data to learn from the samples that have been collected could address both timing issues: filling in between samples and giving an indication of current risk without waiting for lab results.

Site selection

For the initial exploration we decided to focus on the Combe Martin area where we have samples collected at two sites for almost a year by citizen scientists and we also have similar data from the Environment Agency bathing water quality monitoring over the swimming season (May-Sept). The southernmost sampling location is in the River Umber, while the other two locations are in the sea, but close to the river outflow.



We looked at the river catchment feeding into these sites and selected the upstream sensors for water quality, depth and rainfall:



Preparing the data

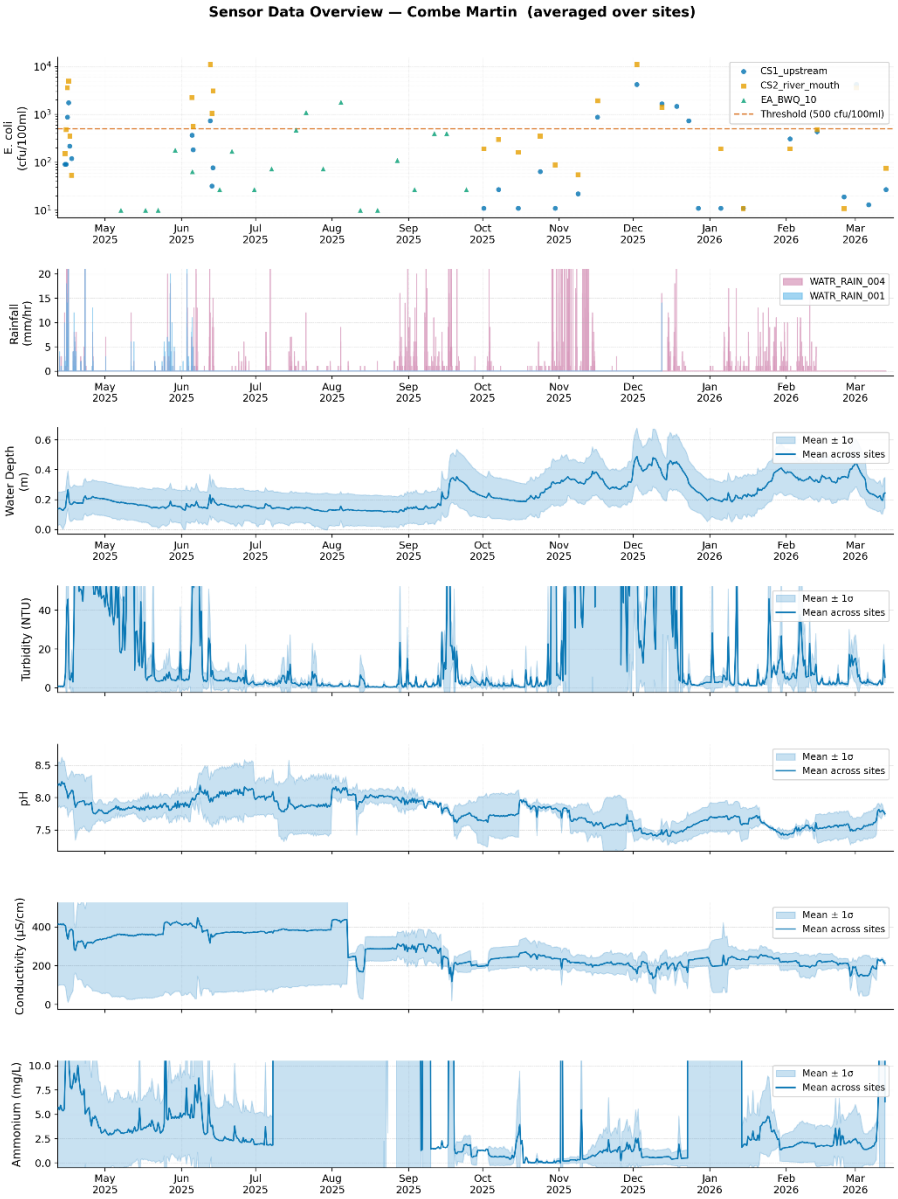
The next challenge was to combine the continuous sensor data with the less frequent samples. The samples will be used as labelled examples to enable the model to find patterns between the inputs (continuous sensor data) and the target we want to predict (E. coli level).

We averaged the sensor readings over time windows before each sample, for example the mean conductivity over the 3 hours before the sample was taken, to build a table of input features for each sample. We also produced datasets in two configurations: with separate sites as individual columns, and combined (averaged across sites for readings of the same type within the time window).

This gives us a table where for each sample we have input features like:

- rainfall (3hr mean and 24 hr max)
- depth (3hr mean and 24 hr range)
- 3hr mean water quality measurements for turbidity, conductivity, pH, ammonium etc.

An overview of the collected data is shown below:



One challenge with in-situ sensor data is reliability. Looking at the ammonium panel for example, there are sustained high plateaus that likely reflect sensor saturation or malfunction rather than real water chemistry. This is a reality of working with continuous environmental sensors and something we need to handle carefully when preparing data for modelling (in this case keeping the high values appears to slightly help model performance – perhaps because they indicated high pollution periods even if the specific numbers were not reliable).

We decided to treat this as a classification problem (predicting a category) rather than regression (predicting a continuous value). Early attempts at regression performed poorly because the very high values are hard to predict (less common in the data and thus we do not have enough information to predict accurately). However, for bathing water quality we are less concerned with the exact value than whether it is above or below a safe threshold and it turns out that this is significantly easier to predict - we used 500 cfu/100ml as the initial threshold – see notes here on the Environment Agency thresholds.

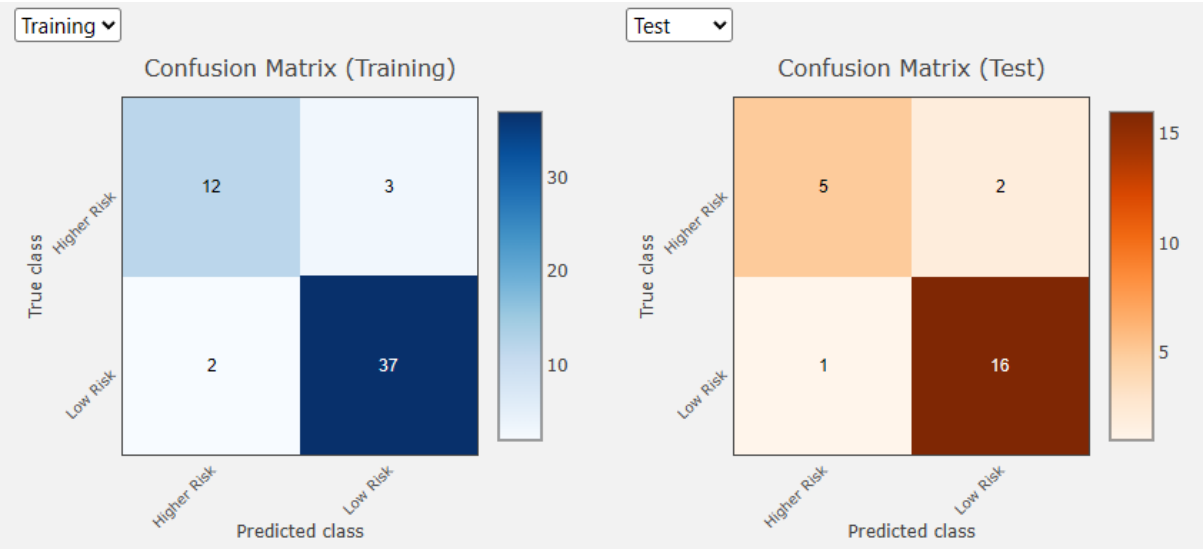
Training a Classifier

When building a model like this we split the data into a training set and a test set - in this case we used a 70/30 split (54 training samples, 24 test). This is important because some machine learning models can fit training data very accurately without generalising well to samples outside this group (overfitting). By holding back a selection at random we get a more realistic evaluation.

Using the 500 cfu/100ml threshold we get a slightly unbalanced dataset with 56 'low-risk' to 22 'higher-risk' samples. This means we have relatively few 'higher-risk' examples, just 15 in the training set and 7 in the test set. This limits the complexity of models worth considering and means that we can't read too much into small changes in the accuracy of the model.

Using a very simple linear Logistic Regression classifier, we found a range of accuracies depending on the test/train split (i.e. if you shuffle and split again) of ~0.8-0.96 (with our sample size this is the difference between misclassifying 1-5 samples out of the 24 test samples).

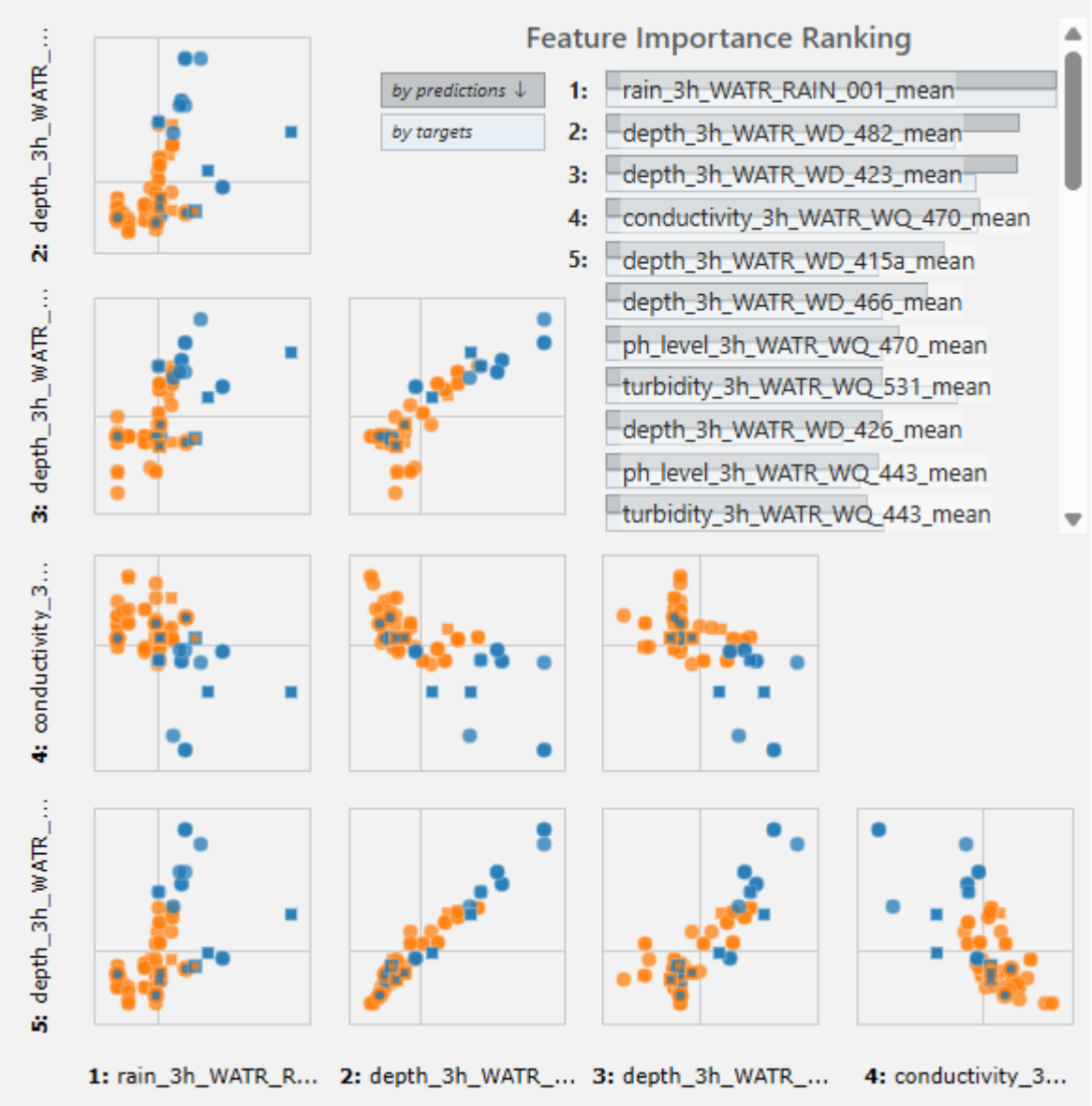
Picking a representative example, we show an example of the confusion matrix and accuracy metrics below. The confusion matrix shows each sample in a quadrant - the correctly predicted higher risk at the top left and correctly predicted low risk bottom right, but also shows the misclassifications: samples predicted higher risk that were actually low, and vice versa:



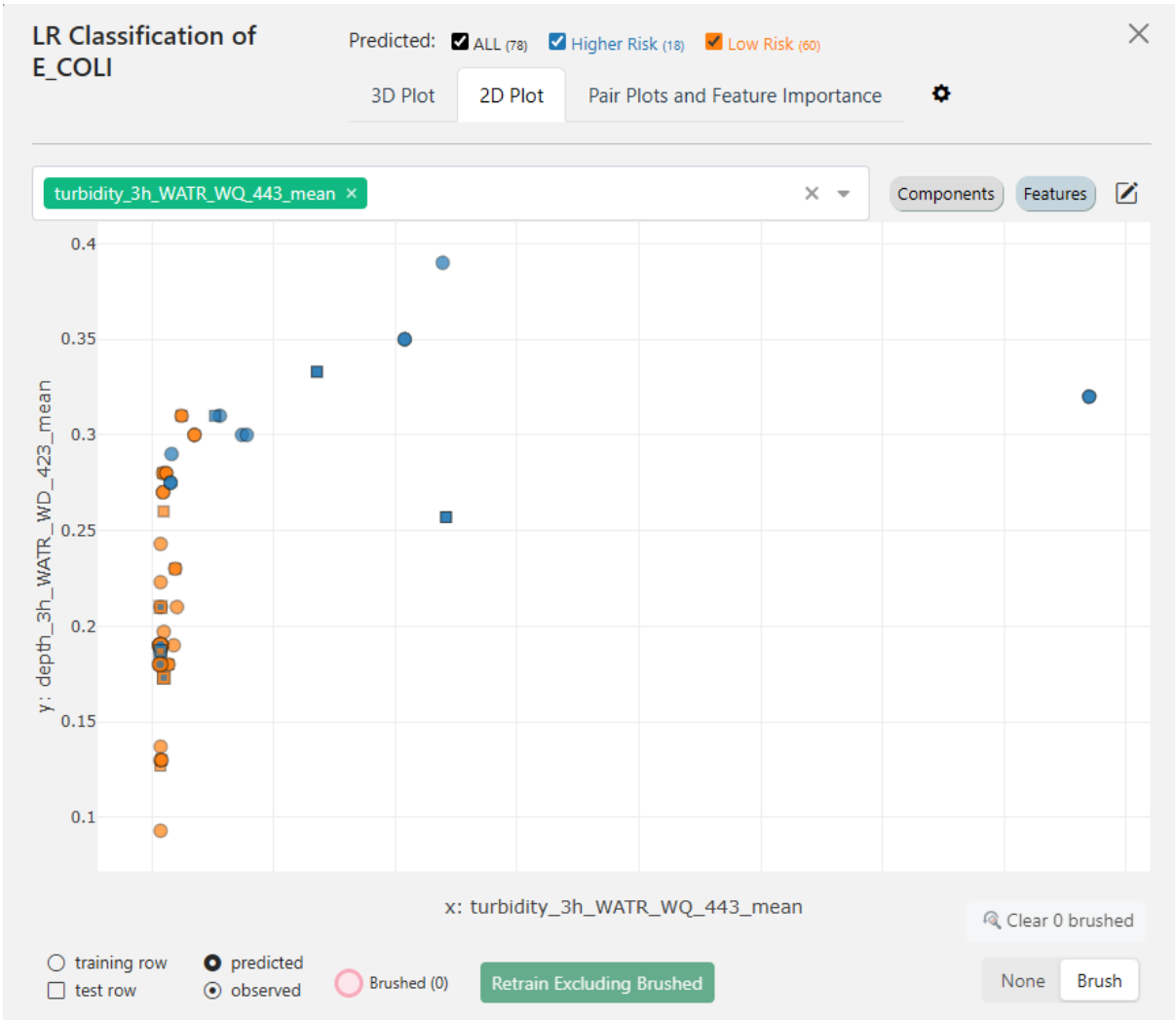
The confusion matrix (test), on the right of the figure, is created using our model and the 24 test samples that our model has not seen before but is trying to predict. The top two quadrants of this matrix contain the higher risk E coli samples, of which there are 7 in total (5 in the top left quadrant + 2 in the top right quadrant). The 5 in the top left quadrant mean our model correctly classified 5 of the 7 higher risk samples. The 2 in the top right quadrant are samples that our model misclassified as low risk, when this sample was in fact higher risk. Similarly, on the bottom row, we can see that there are 17 (1+16) low risk samples, of which our model correctly classified 16 out of the 17 as low risk. The 1 in the bottom left quadrant is a sample our model incorrectly classified as higher risk when it was in fact low risk.

Using the IGI ML tools (https://ml.igiltd.com/) we can investigate which features are important for separating the higher risk samples (blue) from low risk (orange), as shown below. In this case we have treated each sensor as a separate input. The suffix in each label (e.g. WATR_WQ_443) is

the site (location) code. The plot demonstrates that– water depth and turbidity rank highly. This makes sense as both respond to rainfall and surface runoff, which cause bacterial contamination to enter the river, whether from agricultural land, storm overflow discharges or other sources.



Exploring one of these plots in more detail (below) we see a good separation of the higher risk and low risk samples (with some misclassifications).



In further work we intend to explore in greater detail why the samples were misclassified. It could be related to erroneous measurements, or a missing input we have no accounted for, or could not measure. It could also be that the measured values of bacterial load were very close to the 500 cfu/100ml threshold. We also investigated more complex models and got slightly better metrics with a neural network after tuning regularisation (smoothing strength) but would prefer to use a simpler model (Logistic Regression) for a small data set like this.

Limitations and next steps

While the initial results are encouraging, this is an early exploration with a single catchment using a relatively small number of samples. The model has been tested on a random holdout from the full dataset, but not yet validated temporally, for example by training on earlier samples and testing on later ones.

I would be very interested to see how well these patterns generalise as we collect data at more sites. It may be that the relationships are quite localised, in which case we would need a model per site / catchment. Either way, understanding how transferable these models are between catchments is an important question.

There are several directions we would like to explore:

- Including storm overflow data once we have captured the history of when overflows are active over a longer period.
- Experimenting with different risk thresholds to see how sensitive the results are to where we draw the line.
- Trying more complex models as the dataset grows, though with this few samples simpler models remain the safer choice.
- Investigate further the misclassified examples to see if there is a pattern.

There is established work in this area, including the Environment Agency's Pollution Risk Forecasting system which uses rainfall forecasts and other predictions to warn of poor bathing water quality at coastal sites, and the RDMAI Open E. coli Models project which takes a national-scale approach using land cover and environmental data. Our work is complementary to these, exploring what can be learned from in-situ sensors within a specific river catchment.

This work depends on the samples collected by citizen science volunteers coordinated by the North Devon Biosphere. As the dataset grows over time, so will the potential for models like this.