Toxic gas identification - Data challenge report

Kévin Feltrin - ESIEE Paris E4 DSIA - Machine Learning 2 - Professor J-F. Bercher

1. Introduction

This report describes our approach to the Bertin Technologies Data Challenge on toxic gas identification. The task is multi-output regression: predict 23 gas alarm levels in [0,1] from 13 sensor readings (absolute humidity, two resistive sensor banks M4-M7 and M12-M15, and four complementary sensors R, S1-S3). Evaluation uses a weighted RMSE that applies weight 1.2 to samples where the true alarm exceeds 0.5 and weight 1.0 otherwise. All code is in submission_final.py.

2. Exploratory analysis

2.1 Sensor correlations

The pairwise sensor correlation matrix (Fig. 1) reveals three clearly separated blocks. Sensors M4-M7 are all highly correlated ($r > 0.95$), M12-M15 similarly, and the group {R, S1-S3} forms a third looser block (r around 0.3-0.7). Cross-block correlations are weak and often negative.

Sensors within a block are co-located on the device and respond together to the same gas, so their absolute levels are largely redundant. If M4 increases, M5-M7 increase proportionally. The discriminating chemical signal lives in contrasts between sensors, both within and across blocks. This observation directly drove the feature engineering in Section 4.

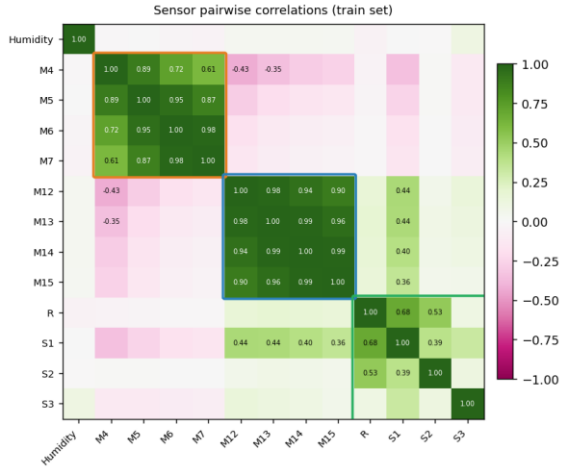


Figure 1: Sensor correlation matrix (train). Three blocks are clearly visible.

2.2 Train-test distribution shift

Table 1 quantifies the humidity gap. The 0.30-0.50 band covers only 1.9% of training but 35.9% of test; the high-humidity regime (>=0.50) is also heavily underrepresented. To measure the global shift, we trained a GradientBoostingClassifier on the 13 raw features to separate train from test samples. The result was AUC = 0.978 under 5-fold CV, confirming near-perfect separability. Looking at individual feature distributions (Fig. 2), Humidity, M4 and M5 show the largest gap between train and test, which is consistent with the challenge instructions warning about humidity and concentration conditions absent from training.

Humidity band	Train	Test
< 0.30	74.1%	21.3%
0.30-0.50	1.9%	35.9%
>= 0.50	24.0%	42.8%

Table 1: Humidity distribution shift.

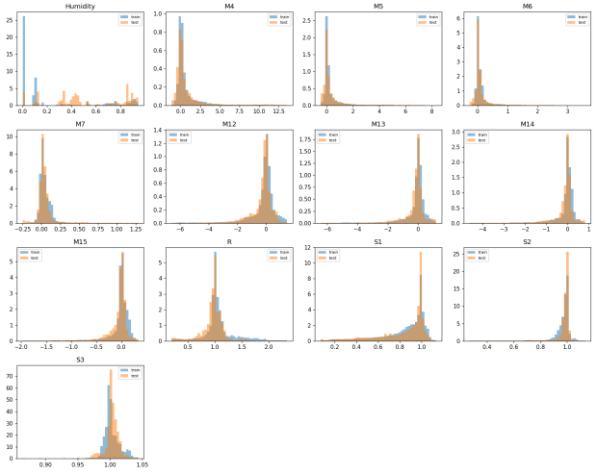


Figure 2: Train vs test sensor distributions. Humidity, M4 and M5 show the largest shift.

2.3 Target structure

The 23 alarm targets are sparse: mean alarm level 0.110, 21.9% non-zero, and only 11.1% fall in the weight-1.2 regime (>=0.5). Figure 3 shows the target correlation matrix. Several class pairs are perfectly correlated because labels combine a gas identity with a semantic descriptor (neurotoxic, vesicant, etc.) and multiple classes fire on the same sample. A multi-output tree with a single shared structure exploits this coupling directly, which independent per-target regressors cannot.

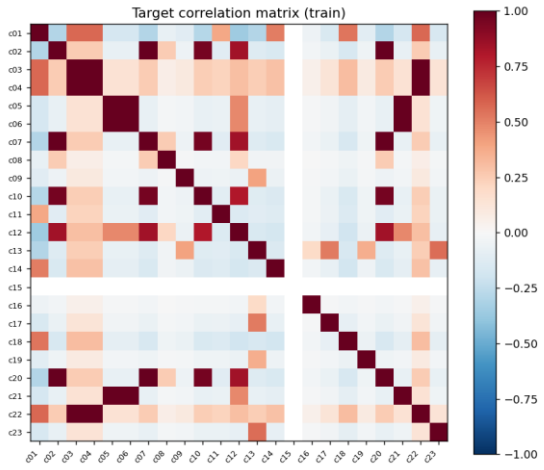


Figure 3: Target correlation matrix. Several pairs are perfectly coincident.

3. Validation

Because train and test are almost perfectly separable (AUC 0.978), a standard random cross-validation fold is drawn entirely from the low-humidity regime and gives an overly optimistic Bertin estimate. We built a humidity-matched holdout: 20% of training samples are drawn so that the humidity histogram matches the test histogram across 10 equal bins (n=29,437 validation samples, training n=173,496).

In practice, even this holdout still underestimates the public Bertin. Every post-hoc correction that improved holdout performance hurt the public score: global additive shifts (+0.005 to +0.015), zero-thresholding at 0.075-0.085, a two-stage classifier/regressor, humidity noise augmentation, and blending with HistGradientBoosting. All were reverted. The remaining discrepancy is likely due to a gas concentration shift mentioned in the challenge instructions that no construction from training data can capture.

4. Feature engineering

Given the block structure in Section 2.1, we augment the 13 raw sensor readings with relative quantities. Since sensors within a block all move together when humidity changes, differences and ratios between them cancel that shared variation. From the 12 chemical sensors we compute: 66 pairwise differences, and 12 ratios of each sensor to humidity. The final feature matrix contains 13 raw columns plus 144 engineered columns, for a total of 157 features.

Table 2 shows an ablation on the humidity holdout comparing different feature subsets (ET-50, same hyperparameters). The 144 engineered features without raw columns achieve the best holdout Bertin (0.0556). The submitted model retains the raw columns alongside the engineered ones, so the tree also has access to absolute sensor levels when deciding which splits to use.

Feature set	#	Bertin
Raw 13 only	13	0.0652
Raw + differences	79	0.0625
Raw + ratios	79	0.0600
Raw + hum. ratios	25	0.0620
144 eng. (no raw)	**144**	**0.0556**

Table 2: Ablation study on humidity holdout (ET-50, d=20, msl=30).

5. Model choice

5.1 Why ExtraTrees

Random Forest and HistGradientBoosting select split thresholds by optimising information gain on the training set. Because the train and test humidity distributions differ strongly, these thresholds are tuned to a humidity range that is largely absent from the test set, which hurts predictions for humidity values outside the training range. ExtraTrees draws split values uniformly at random within each feature range, which avoids this problem. On the humidity holdout, ExtraTrees consistently outperforms both alternatives at matched features (Table 3).

5.2 Hyperparameter choices

For a bagged ensemble, prediction variance decreases as 1/n_trees. Going from 50 to 200 trees reduces variance by a factor of 4 with no change to the bias, which improves the score at only a linear cost in training time (about 12 min on 16 threads). max_depth=20 leaves the trees essentially unconstrained; min_samples_leaf=30 is the real regulariser, ensuring no leaf is fit on fewer than 30 samples. This is important because only 11.1% of alarm labels are above 0.5 (the weight-1.2 regime), so without this floor the model would overfit those rare samples.

Model	Features	Bertin
RF (5 trees, d=7)	raw 13	0.090
RF (5 trees, d=7)	144 eng.	lower
ET-50 (d=20)	144 eng.	0.0556
HGB-200	144 eng.	worse
ET-200 (final)	**157 raw+eng.**	**best**

Table 3: Model comparison on humidity holdout.

6. Results and variable importance

The final model scores 0.14255 on the public leaderboard (leader: 0.1385). Looking at sklearn's impurity-based feature importances on the trained model, sensor-sensor differences contribute the most to splits, followed by sensor-sensor ratios, then sensor/humidity ratios. Raw sensor values contribute the least. This makes sense given the block structure in Section 2.1: when sensors in the same block shift together with humidity, their differences and ratios stay more stable than their absolute values, so the model relies on them to identify the chemical signature.

Among sensor groups, the S-sensors {R, S1-S3} accumulate the highest aggregate importance despite being the smallest group (4 vs. 8 sensors for the two M-banks combined). This suggests they are more chemically selective. The M4-M7 vs M12-M15 cross-block differences are the second most informative family, likely reflecting a spatial asymmetry in gas distribution across the two physical positions on the sensor array.

Per-class difficulty is correlated with alarm rate (Fig. 4). Classes with more high-alarm samples carry the highest metric weight (1.2) and are simultaneously the rarest in training, concentrating both the error penalty and the data scarcity. The shared tree structure partially compensates: when two correlated classes both fire (visible in Fig. 3), the model captures both with the same splits.

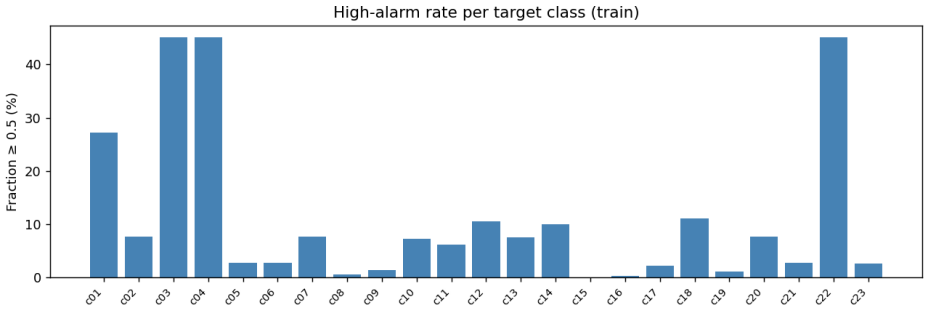


Figure 4: High-alarm rate per target class. Harder classes concentrate the metric penalty.

7. Conclusion

Adding 144 engineered relative features (pairwise differences, ratios, humidity normalisation) to the 13 raw readings reduced holdout Bertin from 0.0652 (raw only) to 0.0556 (engineered only, Table 2), confirming the value of relative quantities over absolute sensor levels. ExtraTrees with uniformly random split thresholds is the right choice under humidity shift because it avoids the threshold bias of greedy methods. Scaling to 200 trees provided a further variance reduction by a factor of 4 at linear cost.

The remaining gap to the leaderboard leader likely comes from two sources: single-seed variance under distribution shift (cross-seed blending would reduce this) and block-level summary features (mean, std, range per sensor block) that the correlation analysis motivates but that were not included before the submission deadline.