Continuous Anomaly Detection and Labeling for the Fermilab Linac: L-CAPE

Jason St. John, Aleksandar Dyulgerov, Beau Harrison, Kyle Hazelwood, Bill Pellico,
Brian Schupbach, Kiyomi Seiya, Brillina Wang, Davion Washington, Michael Wesley
–Fermi National Accelerator Laboratory, Batavia, IL–
Vinay C. Amatya, Milan Jain, Gihan U. Panapitiya, Jan Strube
–Pacific Northwest National Laboratory, Richland, WA–

High-Dimensional Machine States During Unplanned Linac Beam Outages

Data from ~3k Linac devices were analyzed, comparing their time series to the output of a median filter to identify the largest “anomaly” signals. 24 devices led all others, selecting them for further study.

Down-sampling the 24 time series to 1 Hz, normalized data from successive 10-second intervals were reshaped to give 240-dimensional “machine state” vectors.

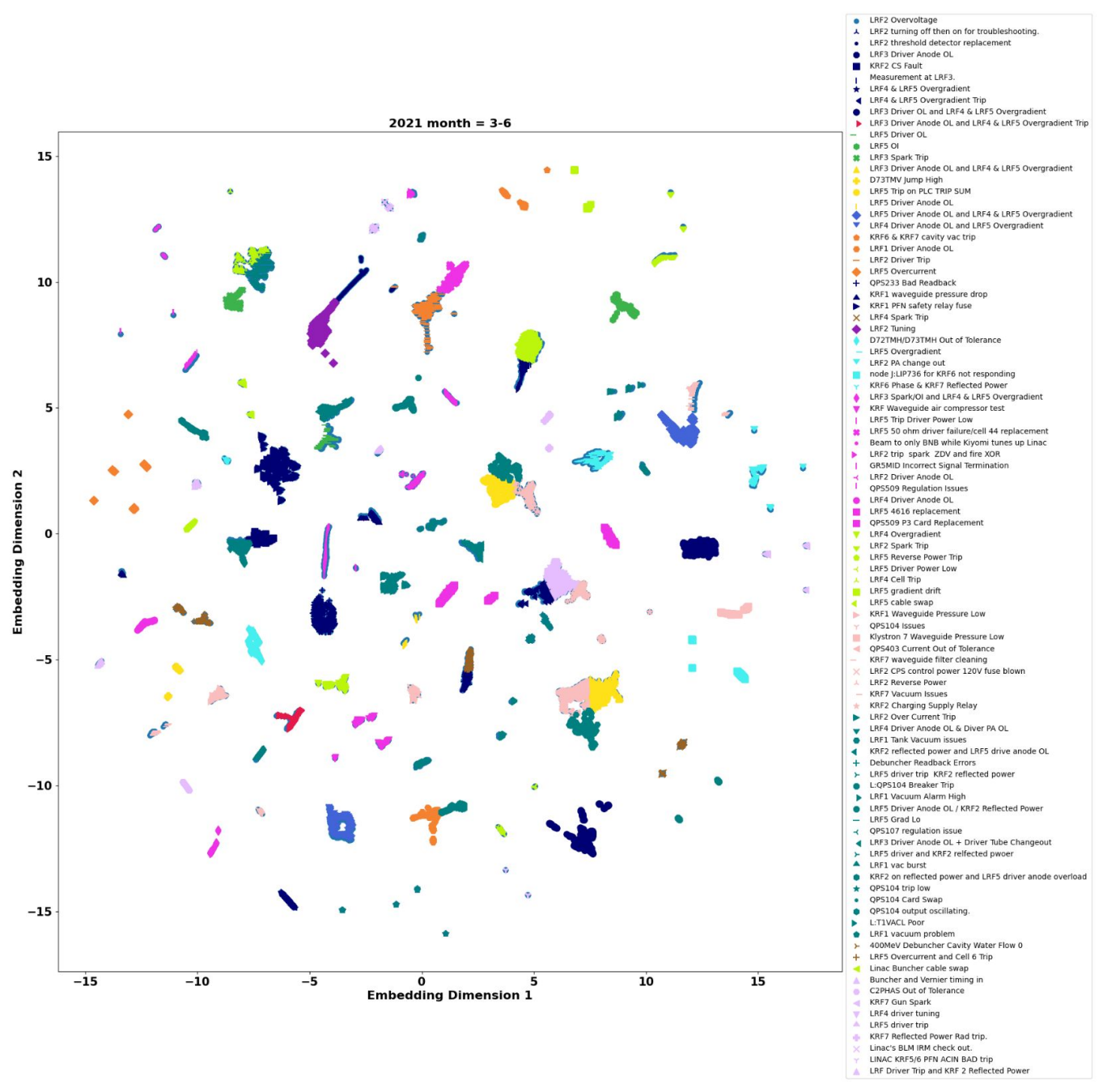

Figure 1. One possible UMAP 2D embedding for the 12,377 state vectors from down times 2021.03-06.

UMAP (Uniform Manifold Approximation and Projection) is a fast-fitting dimensional reduction algorithm which preserves local structure in the low-dimensional embedding. Fig. 1 shows one such UMAP embedding for these outage data using default parameters (random seed, min_dist, and n_neighbors) where human-applied labels convey the reason for the beam outage. Many outages form distinct “islands.”

Live data can be readily visualized in this same space, automatically suggesting labels. A proof-of-concept site has demonstrated this in real time.

UMAP Parameter Optimization

Seeking compact embedding groups, we fit UMAP with 200 random seed values at the values of `min_dist` and `n_neighbors` shown in Fig. 2. Figure of merit taken to be the mean diagonal span (root-sum-squared extent in low-dimensional projection) of each single-label point cloud, population-weighted averaged across labels. Minimized (optimized) at `min_dist=0.2`, `n_neighbors=200`.

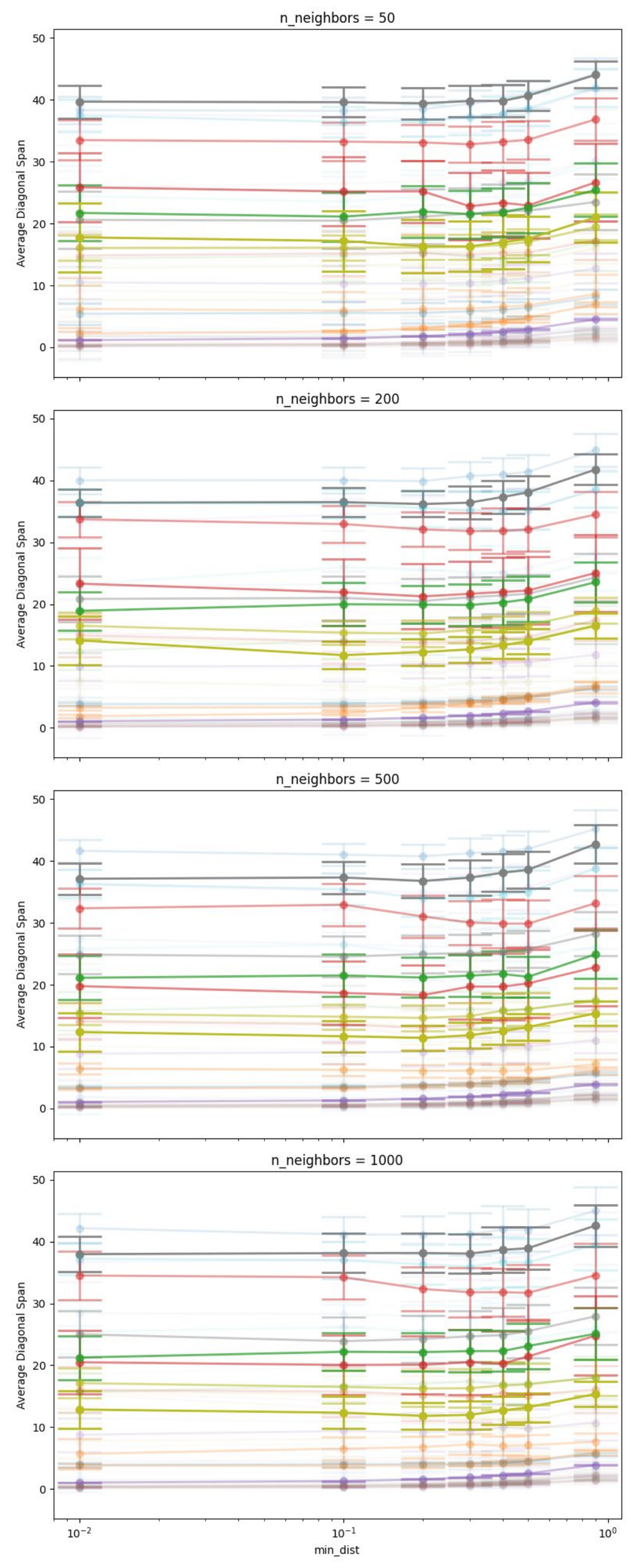

Figure 2. Diagonal span (mean ± stddev) of labeled groups vs. `min_dist` value, for several choices of `n_neighbors`. Opacity scales with label population size.

ML-Assisted Ground-Truth Cleanup

Several label islands overlap neatly. In most cases this shows where human-applied labels were not standardized, and should be merged.

A study of nearest neighbors was conducted to help carry out label-merging systematically: For the nearest point to each label’s data points, what fraction (“affinity score”) bear each label? See heatmap, Fig. 3 for result at default-value random seed.

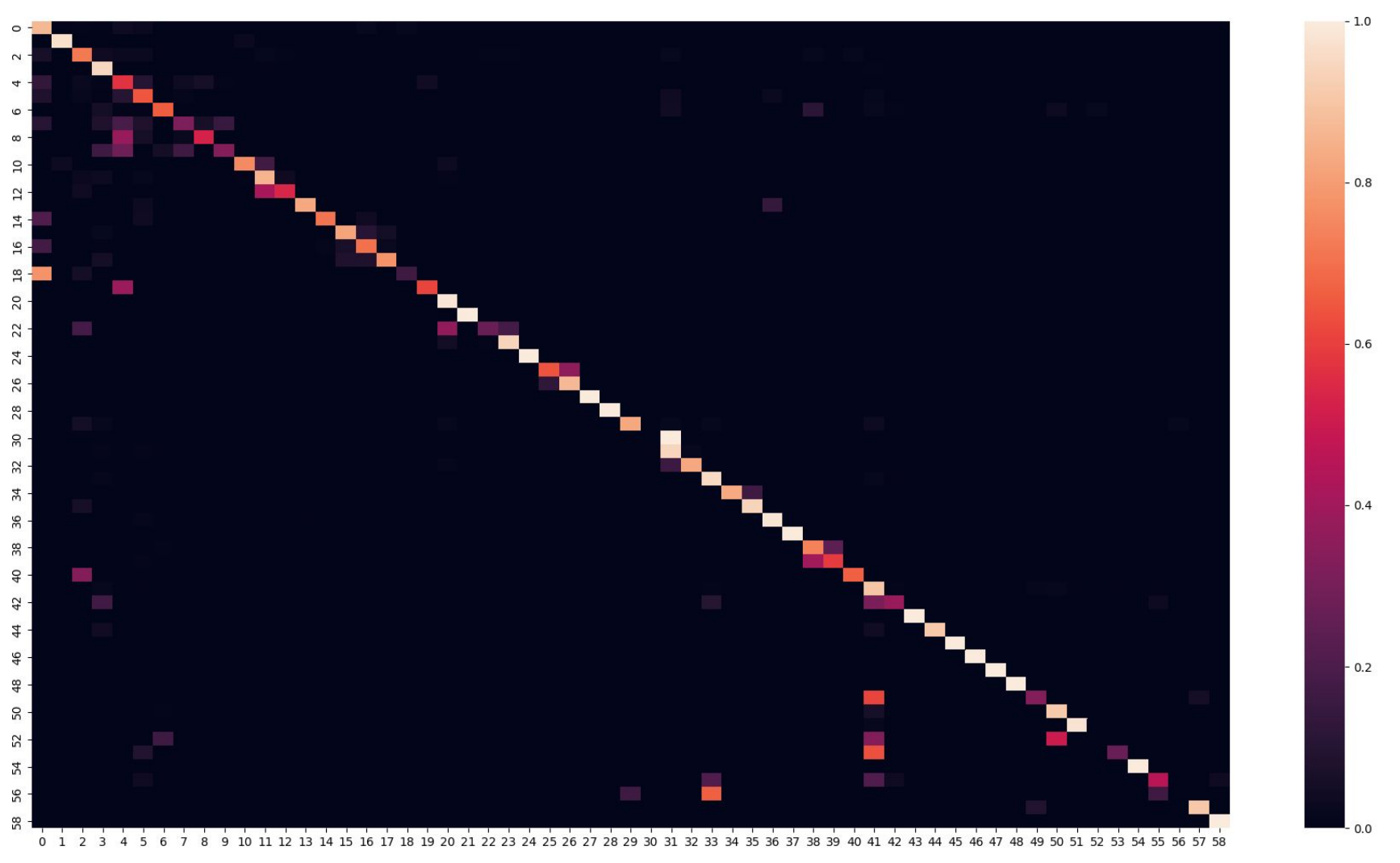

Fig. 3. Heatmap of nearest-neighbors study in the low-dimensional UMAP embedding. Among the off-diagonal “affinity” scores, this study revealed some labels for which a substantial fraction (up to 100%) of the data points’ nearest neighbor is labeled differently. On-diagonal “purity” scores range from 0.0 to 1.0.

For example, all six datapoints of label 30 (“LRF5 Reverse Power Trip”) were nearer to one of the 1231 data points of label 31 (“LRF5 Driver Power Low”), suggesting this is a redundant label. Conducting successive mergers A → B removes 26 of the original 59 labels, merging them into 12 larger populations, removing all inter-label affinities above 0.05, and all self-affinity scores below 0.66.

Affinity scores improve as expected (Table 1) and the population-weighted span figure of merit used in optimization increases, reflecting the higher population of some of the widest-flung labels.

	On-Diagonal	Non-Zero Off-Diagonal
Before	Mean: 0.7357 Median: 0.8276 Stddev: 0.2865	Mean: 0.0736 Median: 0.0183 Stddev: 0.1393
After	Mean: 0.9394 Median: 0.9765 Stddev: 0.0885	Mean: 0.0263 Median: 0.0069 Stddev: 0.0570

Table 1: Statistics of nearest-neighbor “affinity” scores before (upper) and after (lower) one round of label mergers

In a future study, the nearest-neighbor matrix may be recalculated after each merger, which may give slightly different results. Conducting this analysis over many UMAP random seeds will demonstrate the robustness of the merger recommendations to randomness, if any.

Thus the nearest-neighbor analysis of the UMAP projection (1) quantifies spatial clustering of ground-truth labels in low-D space and (2) mechanistically proposes label mergers to improve the human-applied ground-truth labels, improving method scalability.

Future work

Planned refinements: Expanding training data history; Restricting to “onset” datapoint of each outage; Faster intervals and expanded set of devices in machine state.