Supplementary Materials for Association Mapping in Biomedical Time Series via Statistically Significant Shapelet Mining

Christian Bock, Thomas Gumbsch, Michael Moor, Bastian Rieck, Damian Roqueiro and Karsten Borgwardt

These supplementary materials contain additional examples and visualizations that go beyond the scope of the main paper. We first show additional visualizations of the contingency tables for each data set.

S1 Contingency Table Plots

Given a set of statistically significant shapelets obtained from our method, we defined a transformation in the paper that permits us to visualize *all* contingency tables analyzed during the shapelet mining process. In the paper, we only show the plot for the heart rate data set. Here, Figure S1 shows the plot for the respiratory rate data set, while Figure S2 shows the plot for the blood pressure data set. For both plots, we show all contingency tables analyzed by our method in gray, while the statistically significant ones are shown in read.

Notice that these visualizations exhibit a very grid-like structure in the gray points. This is a visual indicator of Tarone's insight, namely that only finitely many values can be obtained for each contingency table. More precisely, given $n_1 + n_0$ different time series, the relations $a_S + b_S = n_1$ and $d_S + c_S = n_0$ imply that there are only $n_1 \cdot n_0$ different contingency tables. Hence, even though the plot has the appearance of representing a smooth structure, this is not true in reality—formally, we are visualizing a finite subset of $\mathbb{Q} \subseteq \mathbb{R}$.

Respiratory Rate The plot shown in Figure S1 has the same characteristics as the one we show in the main manuscript. For this data set, statistically significant shapelets are situated in the first quadrant (upper-right corner). This indicates that their contingency tables satisfy $a_S \gg b_S$ and $c_S \gg d_S$. Interestingly, *no* contingency tables in the third quadrant (lower-left corner) are observed here.

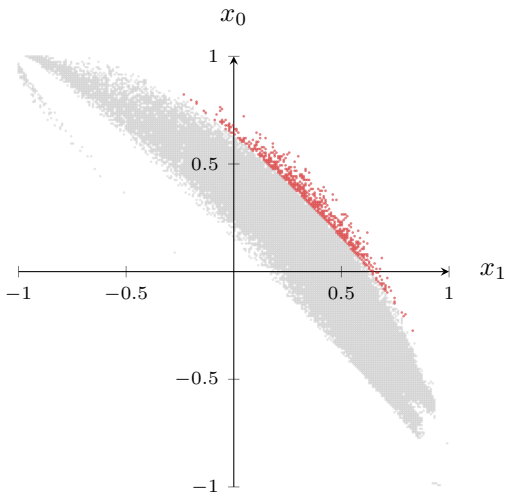

Figure S1: All contingency tables for the respiratory rate shapelets.

Systolic Blood Pressure Figure S2 contains fewer points because, as we outlined in the main manuscript, we used a random sample of the data set due to missing variability. In contrast to the other data sets, this one exhibits shapelets in the lower-left quadrant as well. Due to the selected significance threshold, we can see that *none* of the shapelets in that quadrant is deemed significant.

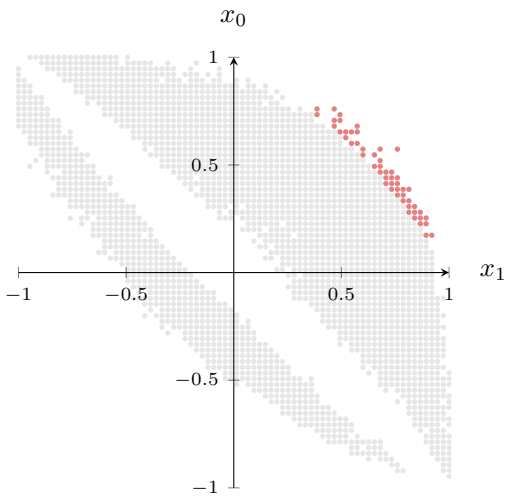

Figure S2: All contingency tables for the blood pressure shapelets.

General Observations We observe that the statistically significant shapelets are always situated at the "border" of the point cloud and form a relatively distinct cluster. In particular, the cluster does *not* contain any gray points (except for the transition to the boundary, where red points have gray neighbors). This is a visual demonstration of the concept of testability to some extent: if one contingency table is deemed testable, its "neighbors" cannot be all untestable. As mentioned in the paper, ideally one would like to have some shapelets in the upper-right corner and the lower-left corner, because this would indicate contingency tables with either $a_S \approx n_1$, $c_S \approx n_0$, $b_S \approx 0$, $d_S \approx 0$ (upper-right corner) or $b_S \approx n_1$, $d_S \approx n_0$, $a_S \approx 0$, $c_S \approx 0$ (lower-left corner). More informally, these contingency tables represent strong splits in the data set. The fact that we do not observe shapelets in the extremal corners sheds some light on the performance that we observed with respect to classification.

S1.1 Properties of Contingency Table Plots

Here, we want build some intuition about the properties of the plot. In general, the plot exhibits symmetry along the line $f(x) = -x$, meaning that if we switch the *rows* of a contingency table (which corresponds to switching the labels in a data set), the resulting point will be *reflected* along the line $f(x) = -x$.

Furthermore, the structure of the plot makes it possible to visually distinguish between data sets with randomized labels and the "real" data set. To demonstrate this, we perform a random permutation of the labels of the heart rate data set. We then compute the contingency table plot of all shapelets analyzed by our method. Note that, as expected, we do not obtain *any* statistically significant shapelets. Figure S3 shows the analyzed shapelets with permuted labels (blue), next to the shapelets with original labels (gray).

Note that the permuted labels exhibit more symmetry because the label of shapelet is decided at random. Furthermore, we can see that the contingency tables are closer to the origin. This hints at their lack of a strong association with one class.

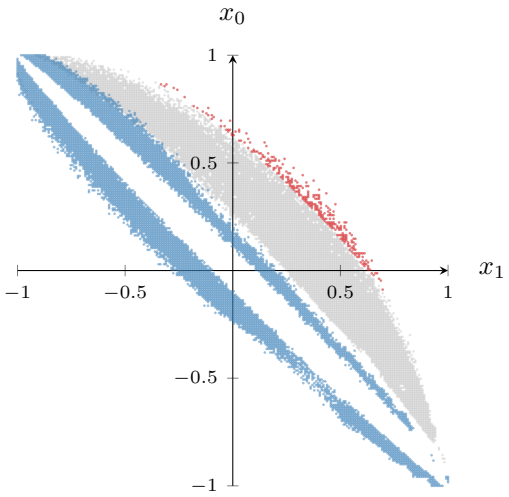

Figure S3: All contingency tables for the heart rate shapelets with original labels (gray) and permuted labels (blue). The statistically significant shapelets detected by our method–using the true labels–are shown in red.