OLET5610 Report

A Preprint

Trent Henderson

then6675@uni.sydney.edu.au

June 20, 2022

1 Method

Time-series classification algorithms are typically benchmarked using the Time Series Classification Repository, which contains 128 univariate (i.e., values sampled uniformly in time) time-series classification datasets/problems and 30 multivariate (i.e., for any time t, $\mathbf{y}_t = (y_{1t}, \ldots, y_{nt})$ describes n realisations at time t) datasets (Anthony Bagnall and Keogh 2022). Here, we focus on the univariate setting. Many algorithms developed across the sciences have been evaluated across the standardised univariate datasets provided in the repository, with black-box algorithms typically outperforming competitors (Bagnall et al. 2017). Given the growing need for algorithmic transparency in the sciences and industry, the present research aims to explore the feasibility of a new approach using one of the 128 datasets as a test case: the "Wafer" dataset (Olszewski, Maxion, and Siewiorek 2001). Specifically, this work aims to determine if using informative summary statistics (known as "features"), can be used to construct a high performing classification procedure that rivals or outperforms existing approaches (Fulcher, Little, and Jones 2013; Fulcher 2017, 2018). Examples of time-series features include properties of the time-series distribution of values, autocorrelation structure, entropy, model-fit statistics, nonlinear time-series analysis, stationarity, and many others (Fulcher and Jones 2014). At its core, feature-based time-series analysis reduces an time series $\times$ time matrix to a time series $\times$ feature matrix, which can then be used for statistical learning (such as classification), using the values of time-series features as inputs.

1.1 Dataset

The "Wafer" dataset in the Time Series Classification Repository contains a collection of inline process control measurements that were recorded from a range of sensors during the manufacturing and processing of silicon wafers for semiconductor fabrication (Olszewski, Maxion, and Siewiorek 2001). Each unique time series in the Wafer dataset represents the measurements of one sensor during the processing of one wafer by one tool. Labels for two classes are provided in the data: *Normal* and *Abnormal*[1]. The goal of this problem is to predict class membership from the time-series values. The dataset contains a pre-designated train-test split, with 1000 samples in the train set (each of length $T = 152$), and 6164 samples in the test set (each of length $T = 152$).

1.2 Algorithmic Approach

The approach in this work contains three stages: (i) extraction of time-series features for each unique time series; (ii) dimensionality reduction using principal components analysis (PCA) to obtain a reduced set of informative vectors; and (iii) classification using the top principal components (quantified in terms of variance explained) as inputs into a random forest classifier. Features from two open-source feature sets (**catch22** (Lubba et al. 2019; Henderson 2022a) and **Kats** (Facebook Infrastructure Data Science 2021)) will be extracted using the R package "theft" (Henderson 2022b). **catch22** extracts 24 features (as mean and

[1] https://www.timeseriesclassification.com/description.php?Dataset=Wafer

standard deviation are added to the standard 22 features) and **Kats** extracts 40 features. This will produce a resulting time series × feature matrix of size 7164 × 64. Given the existence of within-set redundancy (i.e., high absolute correlations between features in a set) observed particularly for **Kats** relative to **catch22** and between-set redundancy identified in previous work (Henderson and Fulcher 2021), dimensionality reduction through PCA will be applied to substantially reduce the input matrix size for the classification algorithm into a time series × principal component matrix. A threshold of 80% cumulative variance explained will be used to determine the number of principal components to retain. Following the procedure of previous work (Ruiz et al. 2021), the methodology presented here then trains and evaluates the accuracy of the classifier over 30 resamples of train-test splits, where each is seeded for reproducibility, and the first is always the pre-designated train-test split in the data as it comes from the Time Series Classification Repository. This will enable scientific inference of algorithmic performance with uncertainty, and facilitate a direct comparison with the performance of existing algorithms and benchmarks.

Given the existence of the Time Series Classification Repository (Anthony Bagnall and Keogh 2022) which holds the Wafer dataset (whose sole purpose is to facilitate benchmarking of time-series classification algorithms), it is expected that meaningful class separation will be possible. The present research takes a novel approach of chaining time-series feature extraction, dimensionality reduction techniques, and a classification algorithm — an approach that has seen almost no research attention to-date. As such, a primary goal of this work is to understand the performance of this approach relative to current benchmarks.

Given the high within-set redundancy (i.e., high absolute correlations between features in a set) observed in previous work, it is hypothesised that dimensionality reduction techniques will substantially reduce the input matrix size for the classification algorithm from the original time series × feature matrix to a much smaller time series × principal component matrix (Henderson and Fulcher 2021). Further, it is also hypothesised that using the time series × principal component matrix as input to a classification algorithm (i.e., random forest classifier) will not result in a substantial reduction in classification performance compared to using the time series × feature matrix nor to existing benchmarks due to the highly informative nature of time-series features in understanding temporal dynamics (Fulcher, Little, and Jones 2013).

2 Results

Prior to substantative analysis, exploratory data analysis was performed. A sample of three time series from each class (*Normal* and *Abnormal*) is displayed in Figure 1. To the eye, some small, but noticeable differences in temporal dynamics and shape are visible between the classes, which suggests that classification based on temporal properties (i.e, "features") is a feasible approach.

2.1 Dimensionality reduction

Time-series features were computed from two sets (**catch22** and **Kats**) on all time series, which produced a time series × feature matrix of size 7164 × 64. This matrix is large, which increases computation time for classification algorithms and increases complexity of model interpretation and evaluation. To reduce this complexity, a principal components analysis was performed. The dataset meets the assumptions of PCA, as multicollinearity between time-series features was a reason for conducting the analysis, and there are sufficient samples to adequately perform PCA as the $7164 : 64$ *samples* : *variables* ratio exceeds the $20 : 1$ recommendation (Osborne and Costello 2019).

Figure 2 plots a visual summary of the PCA. Panel **(A)** plots the percentage of variance explained for the top eight principal components (PC) which together explain 80% of the variance. $PC1$ explains 37.5% of the variance, but there is a steep drop off following this component, with $PC2$ explaining 15.6% of the variance. Panel **(B)** plots cumulative variance explained for the eight retained PCs. The first four PCs explain two-thirds of the variance in the dataset. Panel **(C)** plots the eigenvalues of the eight PCs. All eight PCs exceed the $\lambda = 1$ threshold for the Kaiser criterion of PC retention (Kaiser 1960).

Loadings for each of the 64 time-series features (variables) on the eight PCs is presented in Figure 3. Patterns are evident across the PCs, such as the strong loading of time-series features associated with properties of the autocorrelation and partial autocorrelation function onto $PC1$, and histogram-based statistics onto $PC3$. This plot gives us interpretable and informative insight into the relative behaviour of the time-series features extracted from the Wafer dataset.

Prior to modelling, the PCA was distilled into an even lower dimensional space of just two dimensions to understand if class differences could be ascertained with just the two PCs which explain the most variance

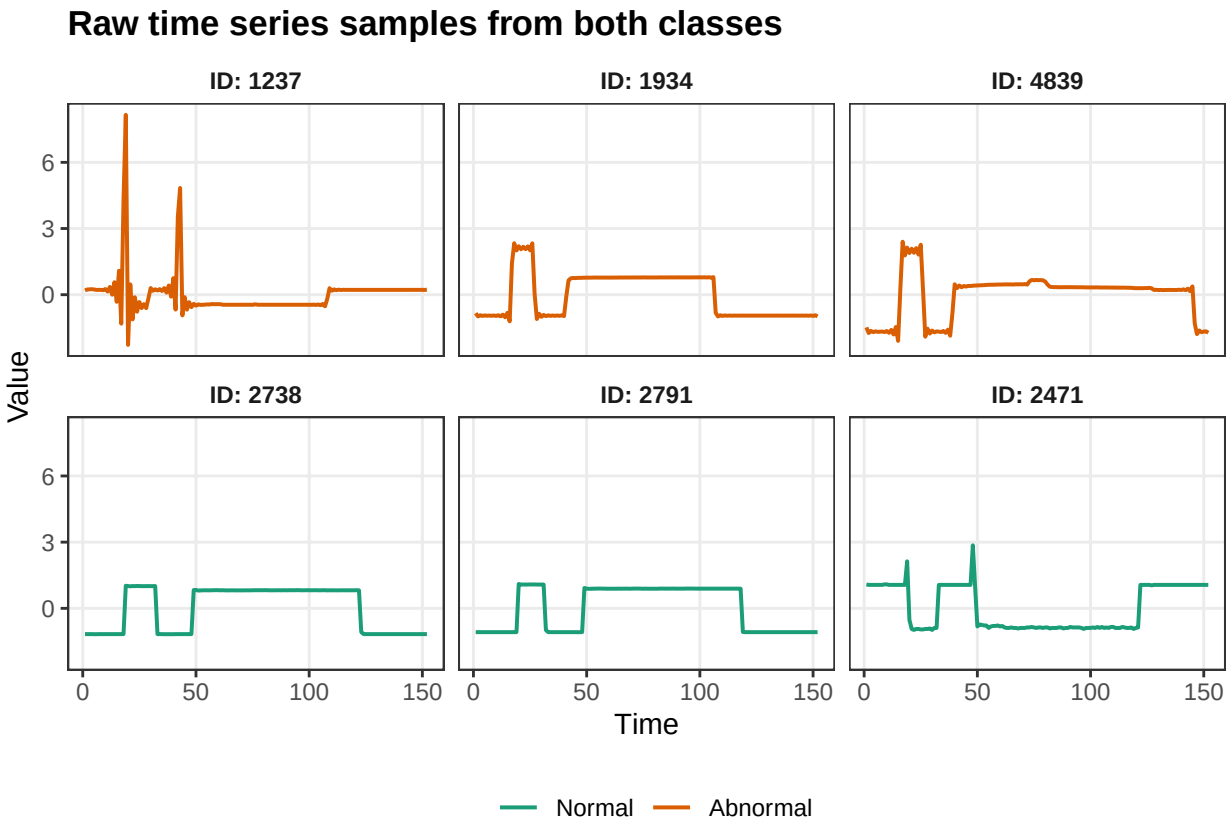

Figure 1: Raw time-series plots of three randomly selected time series from each class (*Normal* and *Abnormal*) in the Wafer dataset. Small differences in temporal dynamics are visible.

in the data (a collective 53.1%). This is displayed in Figure 4). There is considerable overlap between the *Normal* and *Abnormal* classes in the two-dimensional space, suggesting the additional six principal components are likely needed for accurate classification.

2.2 Time-series classification

The time series × principal component matrix (7164 × 8) was passed as input into a random forest classifier using the **caret** package in R (Kuhn 2020) over 30 resamples, where the first sample was the pre-designated train-test split from the Time Series Classification Repository. Mean classification accuracy across the resamples was 99.08%(SD = 0.19%) and is compared to previous benchmarks (Bagnall et al. 2017) in Table 1. The benchmark algorithms include collective of transformation-based ensembles (COTE), shapelet transform (ST), bag of SFA symbols (BOSS), elastic ensemble (EE), dynamic time warping (DTW), time series forest (TSF), time series bag of features (TSBF), learned pattern (LPS), and move-split-merge (MSM). The current approach, while marginally outperformed by the other algorithms, demonstrates classification performance that can be considered to be on-par with more complex, more black-box algorithms, despite only using eight principal components as an input to a random forest classifier. Importantly, no manual hyperparameter tuning was performed, beyond **caret's** basic parameter grid search that it performs over the k-fold cross-validation procedure. It is likely that even stronger performance could be obtained through more detailed hyperparameter tuning and optimisation. Further, relative to the benchmark algorithms presented in Table 1, the current approach is *fast*, executing the entire classification stage (including all 30 resamples, each with 10-fold cross-validation) in under four minutes locally on a laptop with no parallel processing.

To better understand the current approach, we examine the machinery of the random forest models. Understanding relative variable importance can shed a deeper light into random forest models to aid interpretability. Variable importance ranks for each variable (principal components) across the 30 resamples are plotted in Figure 5. $PC4$ (comprised of mostly symbolic features, such as those associated with entropy of small-set probabilities, flat spots, and proportion of magnitudes that exceed a threshold over the standard deviation) is

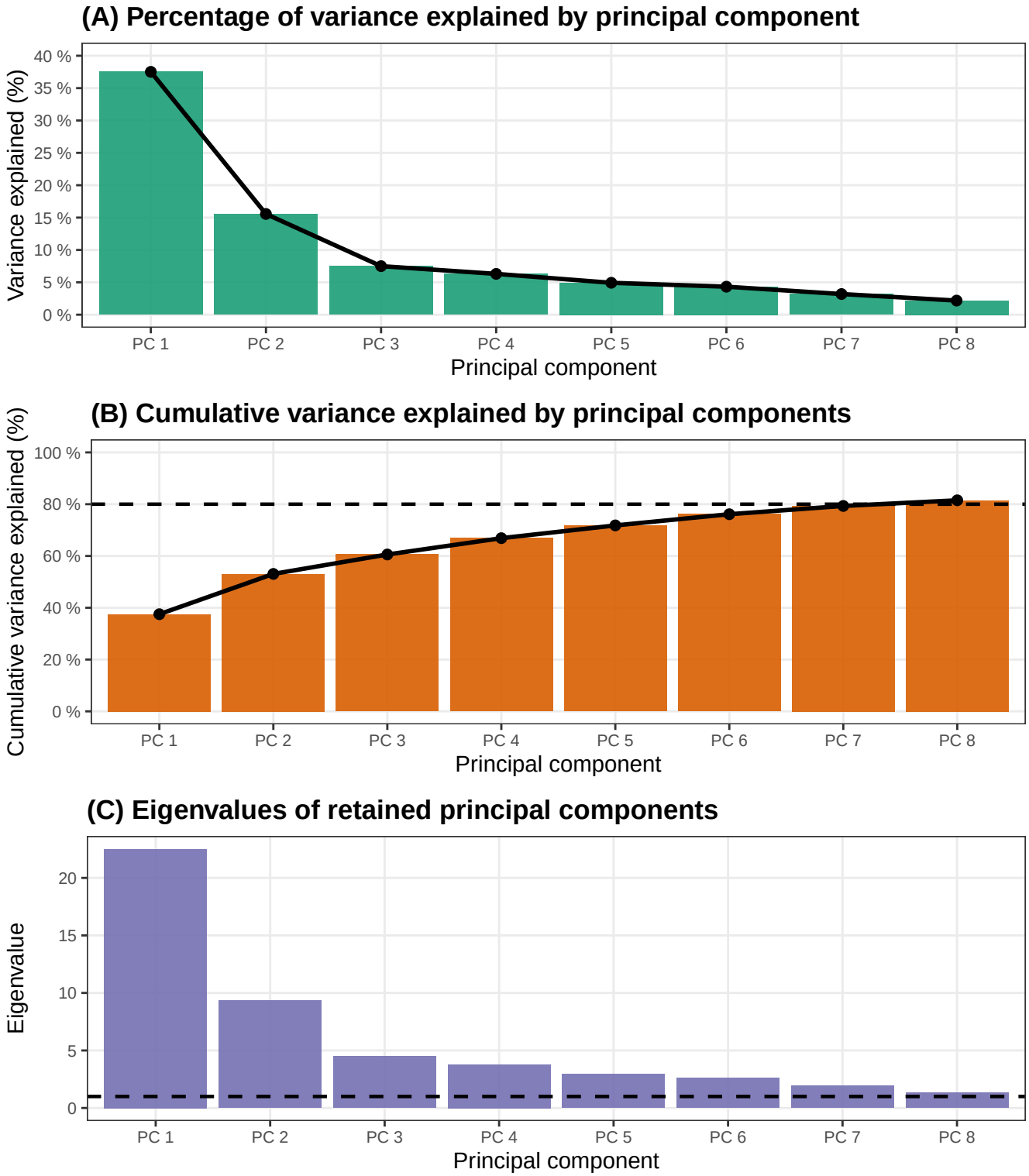

Figure 2: Summary of eight retained principal components. (A) Percentage of variance explained is plotted in descending order for each of the retained principal components. (B) Cumulative variance explained is plotted for each of the retained principal components. An 80% cumulative variance threshold was selected to determine the principal components to retain which returned the eight plotted here (from the original 64). (C) Eigenvalues of the eight retained principal components are plotted in descending order. All retained components also exceed the $\lambda = 1$ cutoff for the Kaiser criterion.

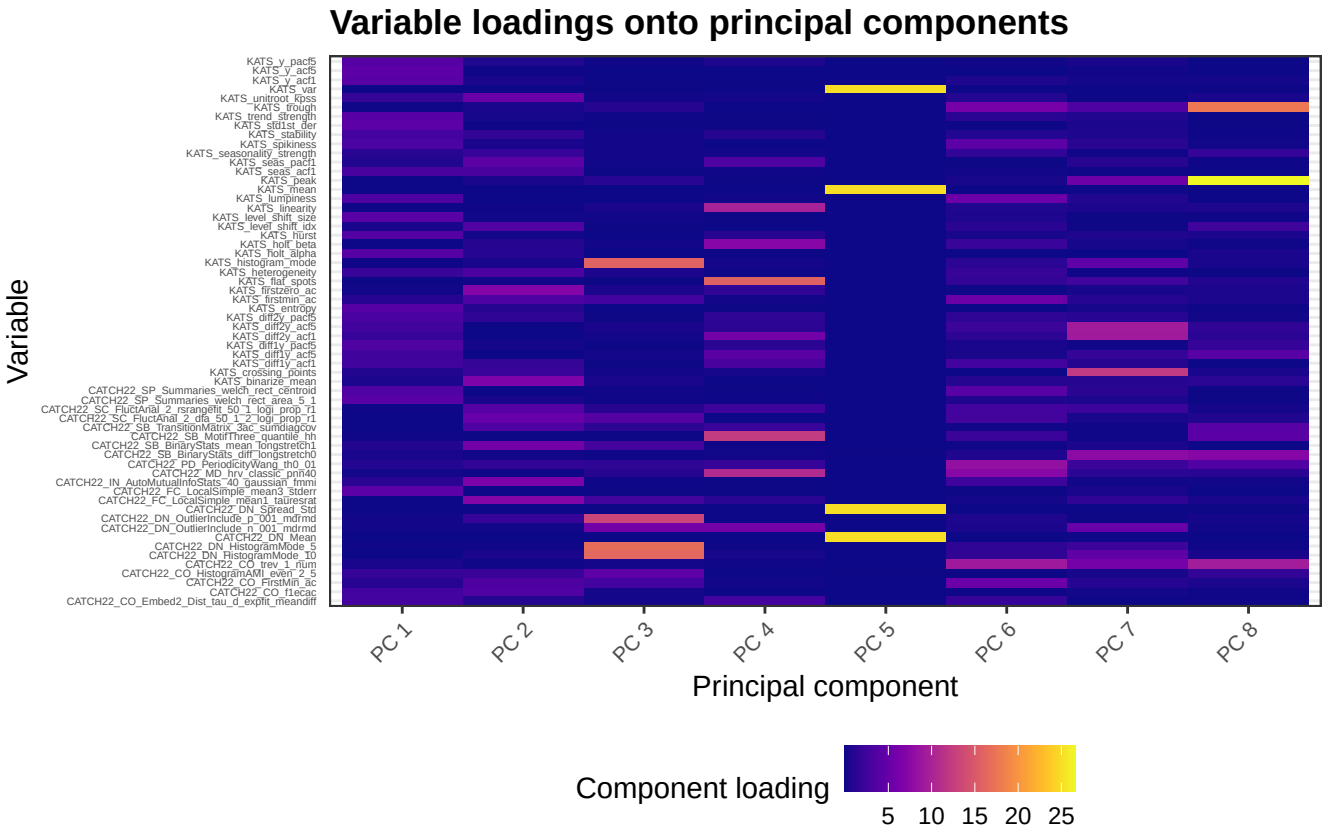

Figure 3: Loadings of each variable onto the eight retained principal components. Relationships between the time-series features are visible, such as loading of features associated with the autocorrelation and partial autocorrelation function onto $PC1$.

Table 1: Comparison of mean classification accuracy results on the Wafer dataset between the current approach and existing benchmarks. Performance differences are minimal between all approaches, with each algorithm achieving 99% accuracy.

Algorithm	Mean Accuracy
ST	100%
COTE	99.9%
BOSS	99.9%
EE	99.7%
TSF	99.7%
DTW	99.6%
TSBF	99.6%
MSM	99.6%
LPS	99.5%
Current approach	99.1%

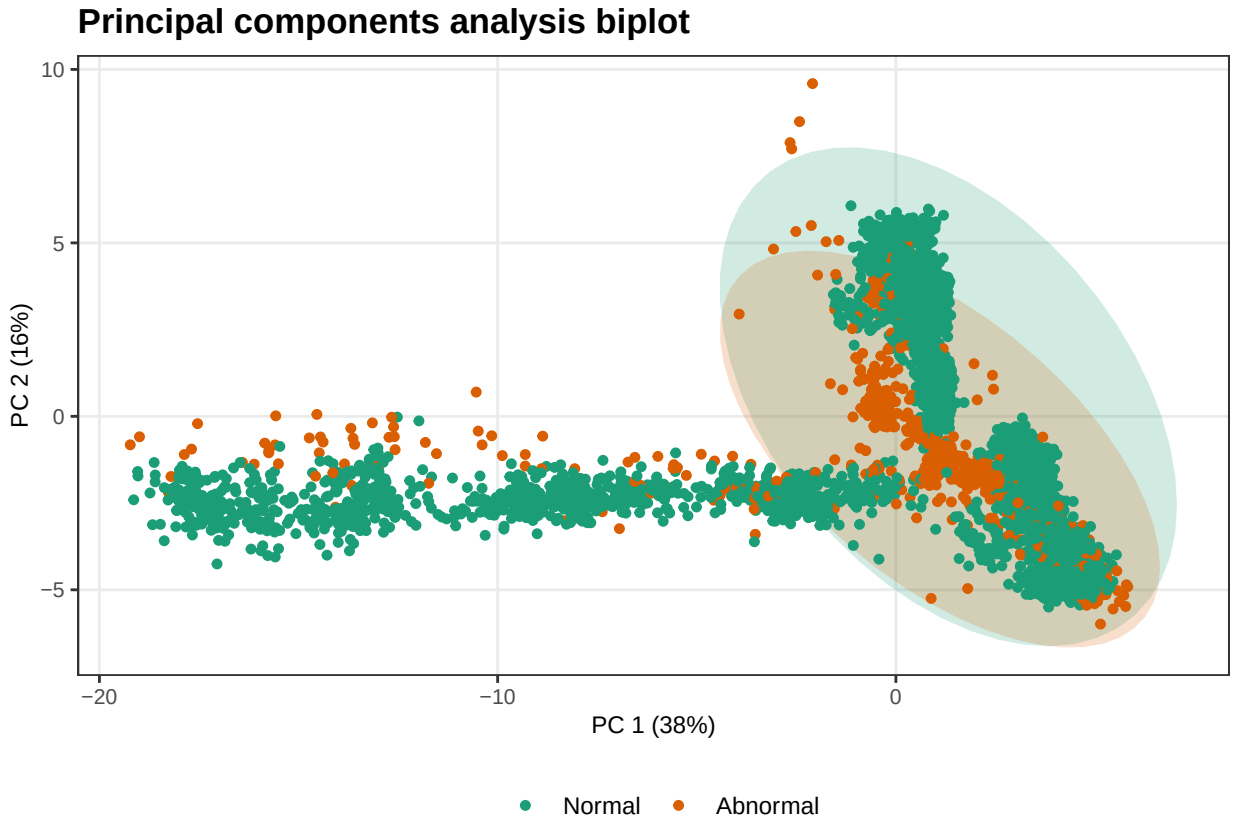

Figure 4: Principal components analysis biplot. The first principal component (positioned along the x-axis) explains 37.5% of the variance in the Wafer dataset. The second principal component (positioned along the y-axis) explains 15.6% of the variance in the Wafer dataset. Class-level covariance is displayed as shaded ellipses.

the most important variable across all 30 resamples for predicting class (*Normal* versus *Abnormal*). The performance of $PC4$ is further demonstrated in Figure 6 which plots the mean $\pm 1SD$ of variable importance values across the 30 resamples. On average, $PC4$ exhibits variable importance of a factor of 5.5 higher than any other variable.

References

Anthony Bagnall, William Vickers, Jason Lines, and Eamonn Keogh. 2022. "The UEA & UCR Time Series Classification Repository." www.timeseriesclassification.com.

Bagnall, Anthony, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh. 2017. "The Great Time Series Classification Bake Off: A Review and Experimental Evaluation of Recent Algorithmic Advances." *Data Mining and Knowledge Discovery* 31 (3): 606–60. `https://doi.org/10.1007/s10618-016-0483-9`.

Facebook Infrastructure Data Science. 2021. "Kats." `https://facebookresearch.github.io/Kats/`.

Fulcher, Ben D. 2017. "Feature-Based Time-Series Analysis." *arXiv:1709.08055 [Cs]*, October. `http://arxiv.org/abs/1709.08055`.

———. 2018. *Feature-Based Time-Series Analysis*. CRC Press.

Fulcher, Ben D., and Nick S. Jones. 2014. "Highly Comparative Feature-Based Time-Series Classification." *IEEE Transactions on Knowledge and Data Engineering* 26 (12): 3026–37. `https://doi.org/10.1109/TKDE.2014.2316504`.

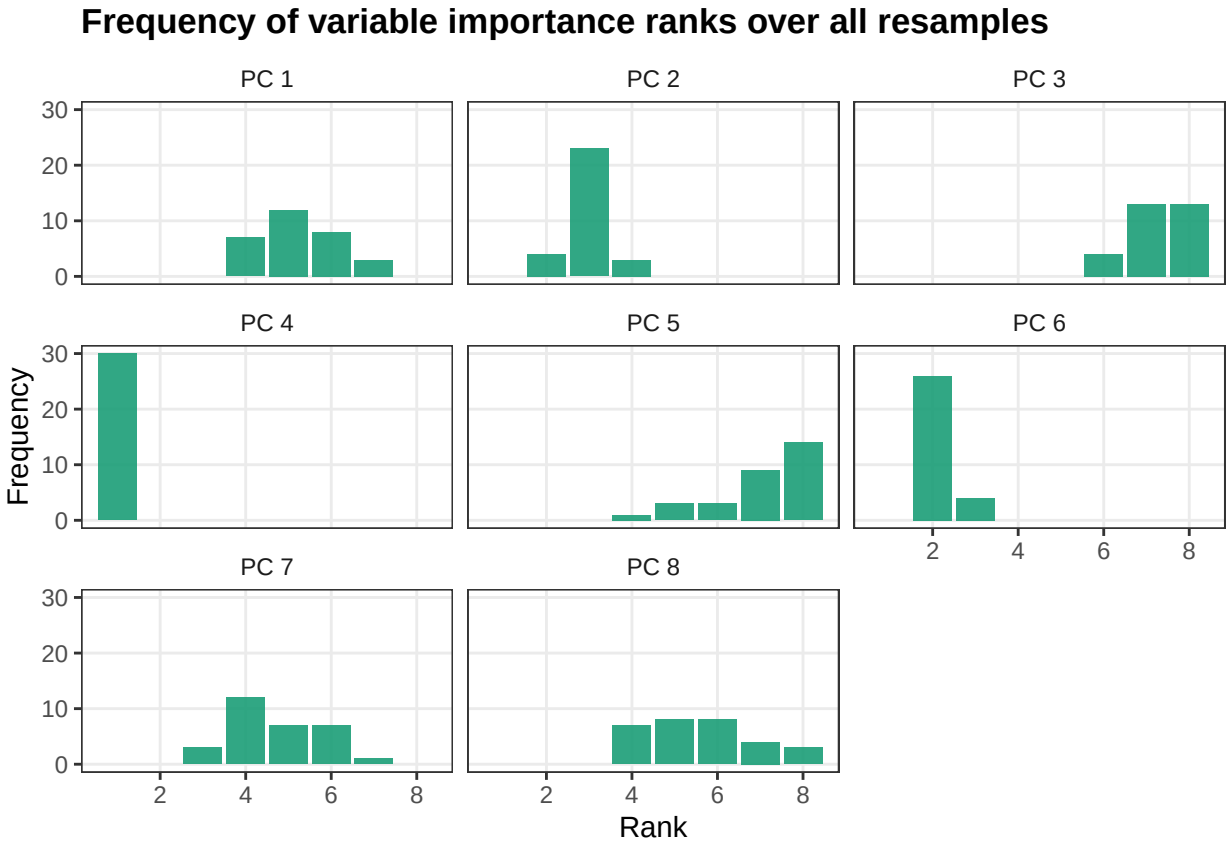

Figure 5: Frequency of ranks over all resamples are plotted for each principal component used as a predictor in the models. *PC*4 is the most important variable across all 30 resamples for predicting class (*Normal* versus *Abnormal*).

Fulcher, Ben D., Max A. Little, and Nick S. Jones. 2013. "Highly Comparative Time-Series Analysis: The Empirical Structure of Time Series and Their Methods." *Journal of The Royal Society Interface* 10 (83): 20130048. https://doi.org/10.1098/rsif.2013.0048.

Henderson, Trent. 2022a. *Rcatch22: Calculation of 22 CAnonical Time-Series CHaracteristics*.

———. 2022b. *Theft: Tools for Handling Extraction of Features from Time Series*. https://hendersontrent.github.io/theft/.

Henderson, Trent, and Ben D. Fulcher. 2021. "An Empirical Evaluation of Time-Series Feature Sets." In *2021 International Conference on Data Mining Workshops (ICDMW)*, 1032–38. https://doi.org/10.1109/ICDMW53433.2021.00134.

Kaiser, Henry F. 1960. "The Application of Electronic Computers to Factor Analysis." *Educational and Psychological Measurement* 20 (1): 141–51. https://doi.org/10.1177/001316446002000116.

Kuhn, Max. 2020. *Caret: Classification and Regression Training*. https://CRAN.R-project.org/package=caret.

Lubba, Carl H., Sarab S. Sethi, Philip Knaute, Simon R. Schultz, Ben D. Fulcher, and Nick S. Jones. 2019. "Catch22: CAnonical Time-series CHaracteristics." *Data Mining and Knowledge Discovery* 33 (6): 1821–52. https://doi.org/10.1007/s10618-019-00647-x.

Olszewski, Robert Thomas, Roy Maxion, and Dan Siewiorek. 2001. "Generalized Feature Extraction for Structural Pattern Recognition in Time-Series Data." PhD thesis, USA: Carnegie Mellon University.

Osborne, Jason, and Anna Costello. 2019. "Sample Size and Subject to Item Ratio in Principal Components Analysis." *Practical Assessment, Research, and Evaluation* 9 (1). https://doi.org/10.7275/ktzq-jq66.

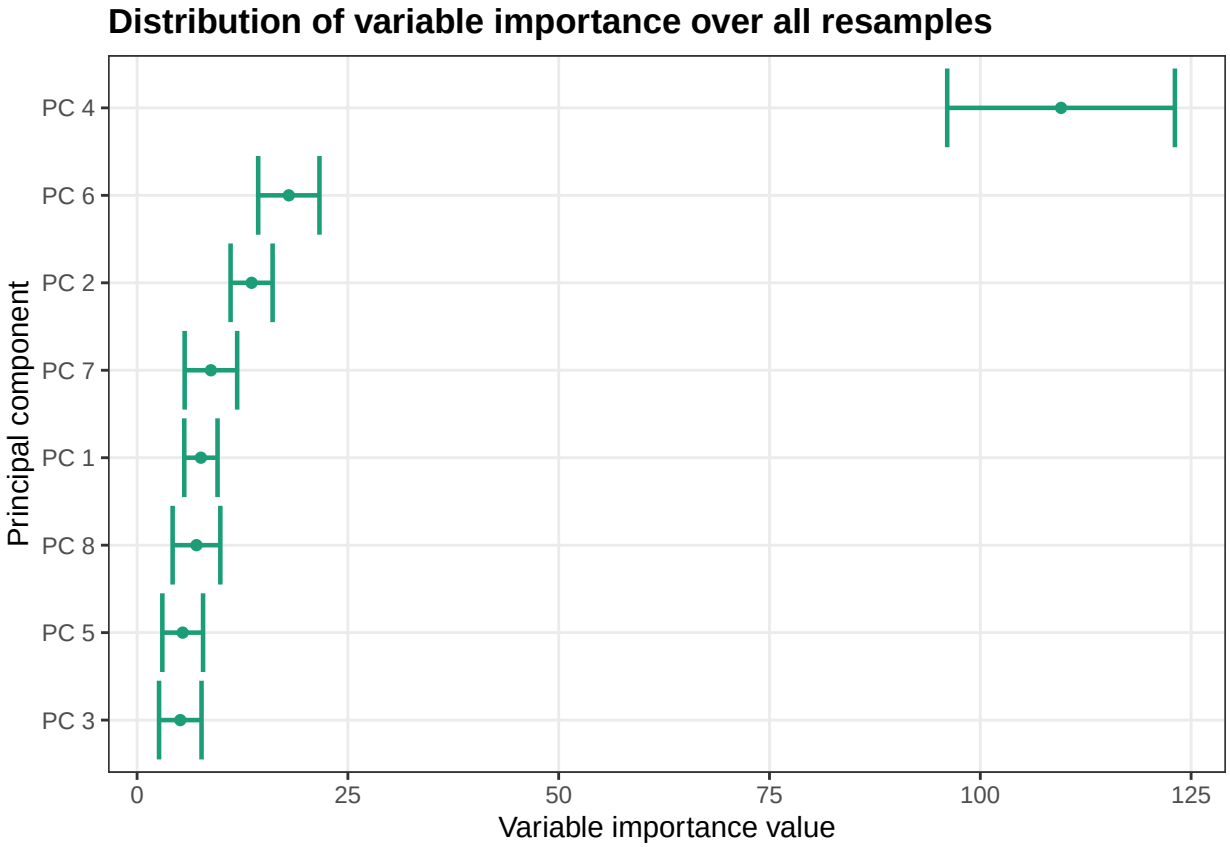

Figure 6: Mean variable importance ± 1SD is plotted for each principal component used as predictors in the models. $PC4$ demonstrates the highest mean variable importance values by a factor of 5.5 over the next highest component ($PC6$).

Ruiz, Alejandro Pasos, Michael Flynn, James Large, Matthew Middlehurst, and Anthony Bagnall. 2021. "The Great Multivariate Time Series Classification Bake Off: A Review and Experimental Evaluation of Recent Algorithmic Advances." *Data Mining and Knowledge Discovery* 35 (2): 401–49. `https://doi.org/10.1007/s10618-020-00727-3`.