Comparison of Clustering Algorithms for Statistical Features of Vibration Data Sets

Philipp Sepin, Jana Kemnitz, Safoura Rezapour Lakani, Daniel Schall

Philipp Sepin
Distributed-Artificial Intelligence-Systems Research Group, Siemens Technology
Vienna University of Technology

TECHNISCHE UNIVERSITÄT WIEN

SIEMENS

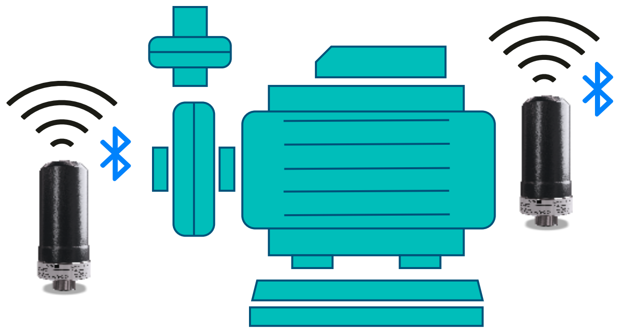
Introduction

Vibration based condition monitoring systems

i) can **accurately identify** different conditions by capturing dynamic features.

ii) enable large scale operations due to **low-cost sensors.**

Research in the field mainly focused on classification or anomaly detection.



Vibration based condition monitoring.

Introduction

Vibration based condition monitoring systems

i) can **accurately identify** different conditions by capturing dynamic features.

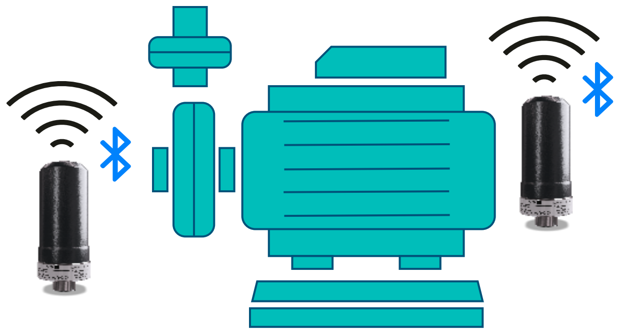
ii) enable large scale operations due to **low-cost sensors.**

Research in the field mainly focused on classification or anomaly detection.

Unsupervised learning methods can prove instrumental as

i) a **preprocessing step** for supervised learning methods.

ii) a stand-alone method when **dealing with missing labels**.



Vibration based condition monitoring.

Challenges

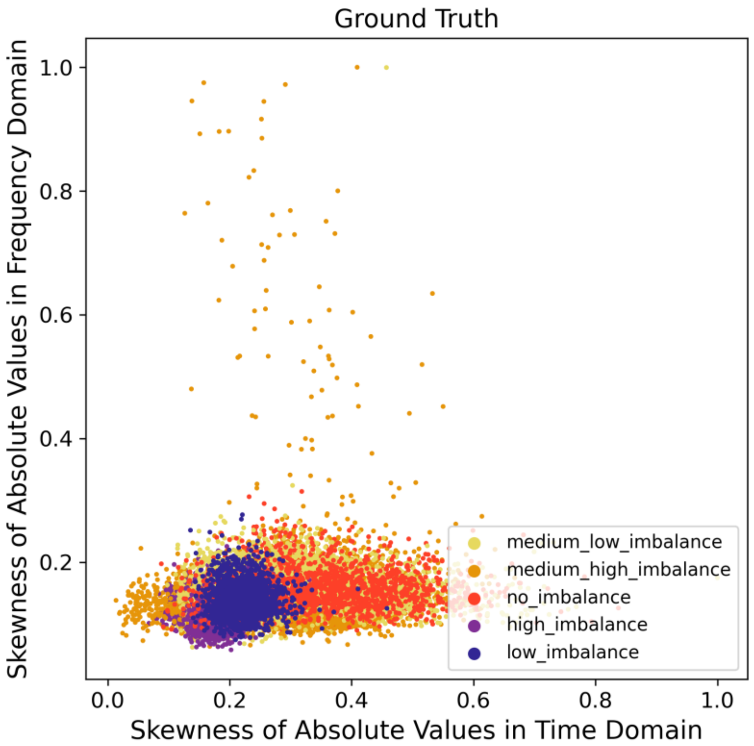
i) Modeling of the feature space to allow for **unsupervised separation of conditions.**

ii) Particularly difficult in the case of industrial data, which is **often inseparable.**

Challenges

i) Modeling of the feature space to allow for **unsupervised separation of conditions.**

ii) Particularly difficult in the case of industrial data, which is **often inseparable.**

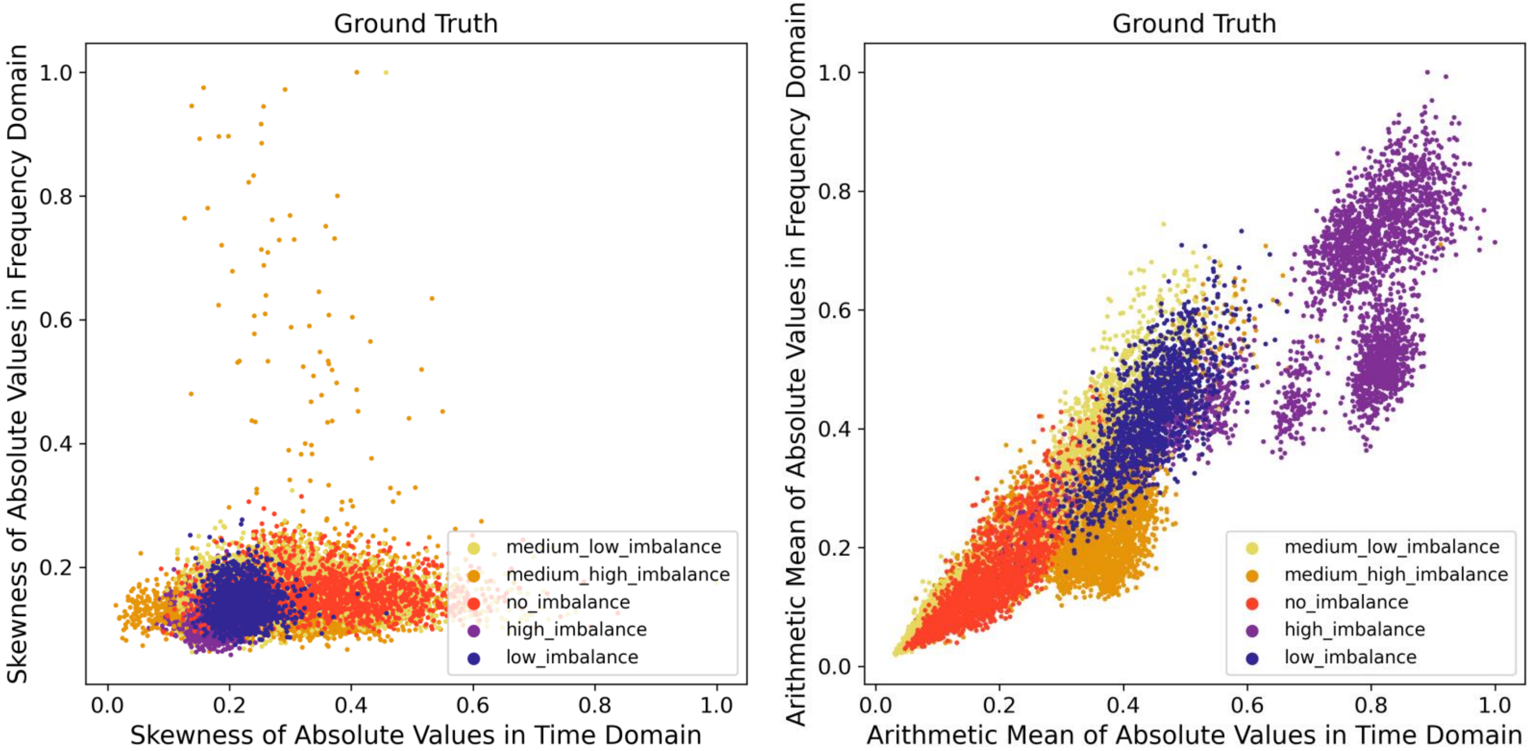


Comparison of ground truth in different feature spaces.

Challenges

i) Modeling of the feature space to allow for **unsupervised separation of conditions.**

ii) Particularly difficult in the case of industrial data, which is **often inseparable.**



Comparison of ground truth in different feature spaces.

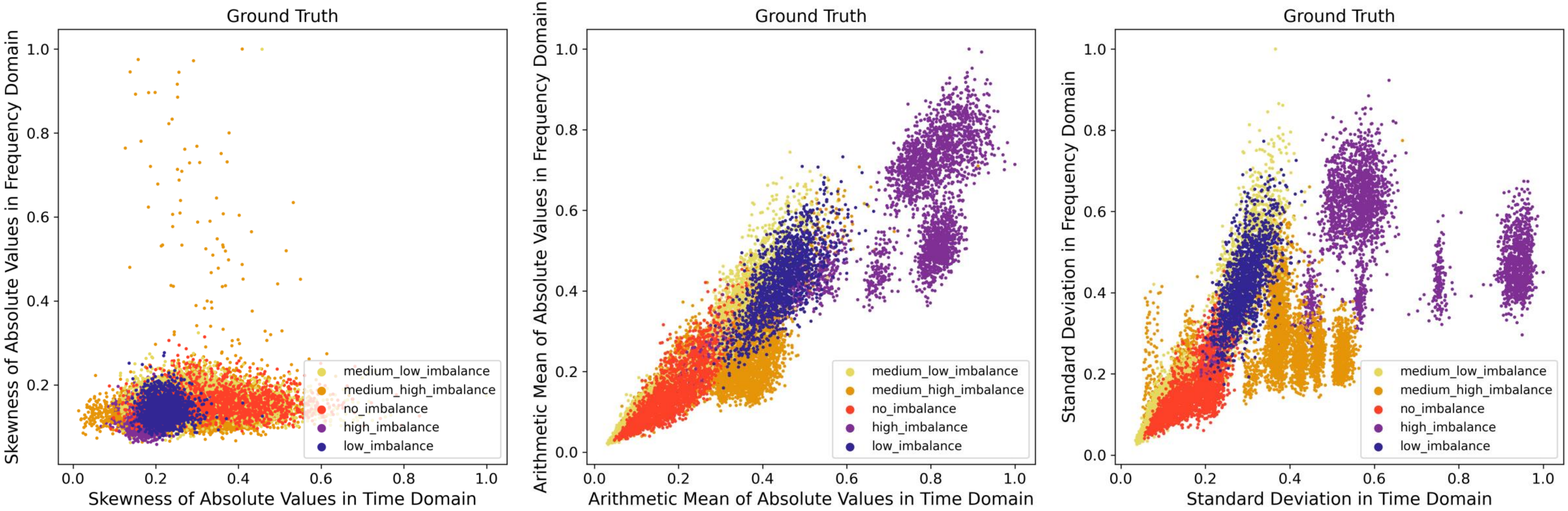
TECHNISCHE UNIVERSITÄT WIEN

SIEMENS

Challenges

i) Modeling of the feature space to allow for **unsupervised separation of conditions.**

ii) Particularly difficult in the case of industrial data, which is **often inseparable.**



Comparison of ground truth in different feature spaces.

Introduction

There is little research on clustering approaches in vibration data, and the solutions are **often optimized for a singe data set**.

TECHNISCHE UNIVERSITÄT WIEN

SIEMENS

Introduction

There is little research on clustering approaches in vibration data, and the solutions are **often optimized for a singe data set**.

We provide a fundamental analysis of

i) **feature extraction and selection methods**.

ii) **clustering algorithms**.

iii) **validated over several data sets**.

We aimed to answer the following questions.

TECHNISCHE UNIVERSITÄT WIEN

SIEMENS

Introduction

There is little research on clustering approaches in vibration data, and the solutions are **often optimized for a singe data set**.

We provide a fundamental analysis of

i) **feature extraction and selection methods**.

ii) **clustering algorithms**.

iii) **validated over several data sets**.

We aimed to answer the following questions.

i) Which **combinations of statistical features and clustering algorithms** perform best for multiple data sets?

Introduction

There is little research on clustering approaches in vibration data, and the solutions are **often optimized for a singe data set**.

We provide a fundamental analysis of

i) **feature extraction and selection methods**.

ii) **clustering algorithms**.

iii) **validated over several data sets**.

We aimed to answer the following questions.

i) Which **combinations of statistical features and clustering algorithms** perform best for multiple data sets?

ii) Does the performance of statistical feature and clustering algorithm combinations **generalize for arbitrary data sets**?

Introduction

There is little research on clustering approaches in vibration data, and the solutions are **often optimized for a singe data set**.

We provide a fundamental analysis of

i) **feature extraction and selection methods**.

ii) **clustering algorithms**.

iii) **validated over several data sets**.

We aimed to answer the following questions.

i) Which **combinations of statistical features and clustering algorithms** perform best for multiple data sets?

ii) Does the performance of statistical feature and clustering algorithm combinations **generalize for arbitrary data sets**?

iii) Can the **combination of several different features** improve the performance of the clustering?

TECHNISCHE UNIVERSITÄT WIEN

SIEMENS

Introduction

There is little research on clustering approaches in vibration data, and the solutions are **often optimized for a singe data set**.

We provide a fundamental analysis of

i) **feature extraction and selection methods**.

ii) **clustering algorithms**.

iii) **validated over several data sets**.

We aimed to answer the following questions.

i) Which **combinations of statistical features and clustering algorithms** perform best for multiple data sets?

ii) Does the performance of statistical feature and clustering algorithm combinations **generalize for arbitrary data sets**?

iii) Can the **combination of several different features** improve the performance of the clustering?

iv) Can principal component analysis (**PCA**) improve the performance of the clustering by selecting the most representative features?

TECHNISCHE UNIVERSITÄT WIEN

SIEMENS

Introduction

There is little research on clustering approaches in vibration data, and the solutions are **often optimized for a singe data set**.

We provide a fundamental analysis of

i) **feature extraction and selection methods**.

ii) **clustering algorithms**.

iii) **validated over several data sets**.

We aimed to answer the following questions.

i) Which **combinations of statistical features and clustering algorithms** perform best for multiple data sets?

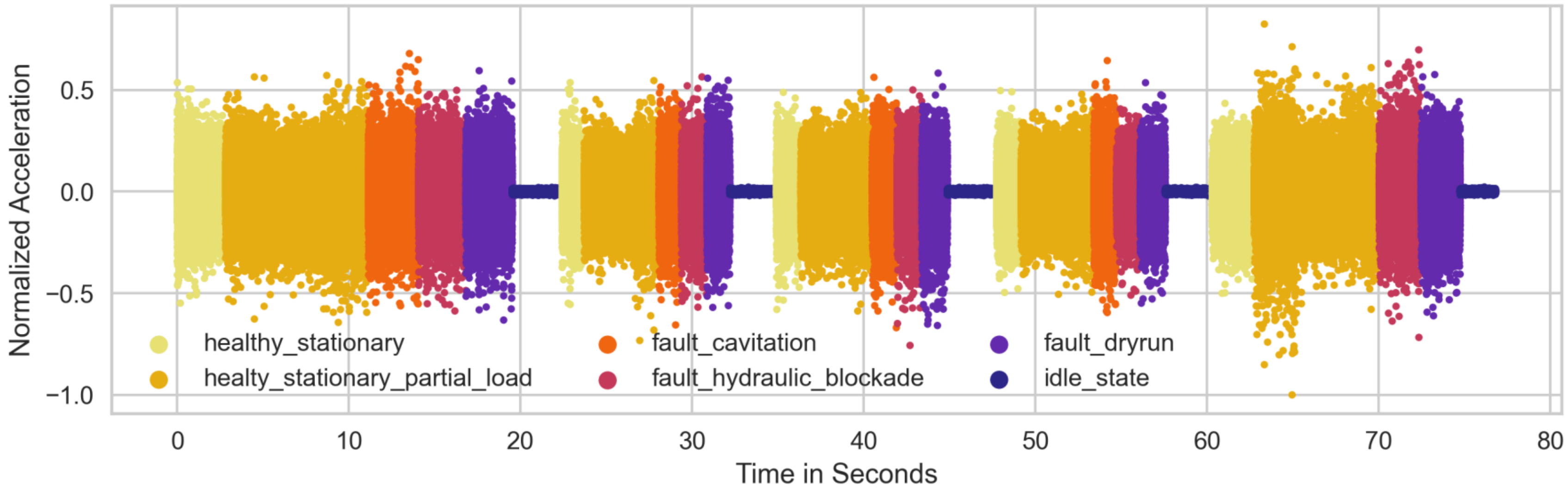
ii) Does the performance of statistical feature and clustering algorithm combinations **generalize for arbitrary data sets**?

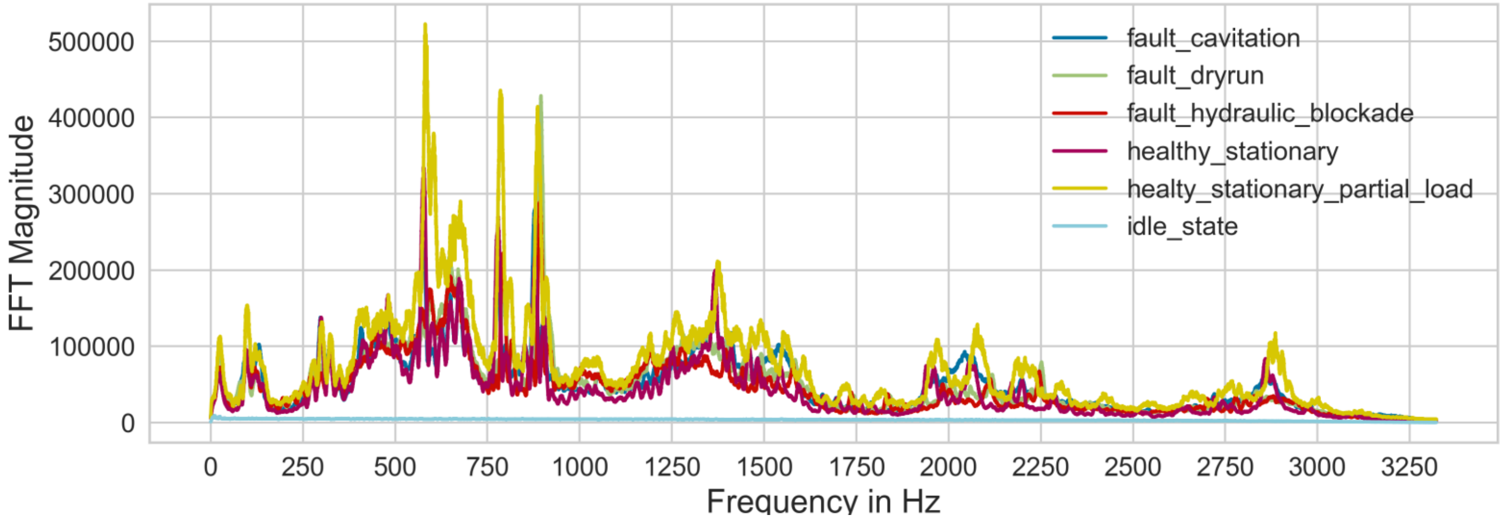
iii) Can the **combination of several different features** improve the performance of the clustering?

iv) Can principal component analysis (**PCA**) improve the performance of the clustering by selecting the most representative features?

v) How does the **specified number of clusters** affect the performance of the clustering?

Data Set 1





Data Set 1 in time and frequency domain.

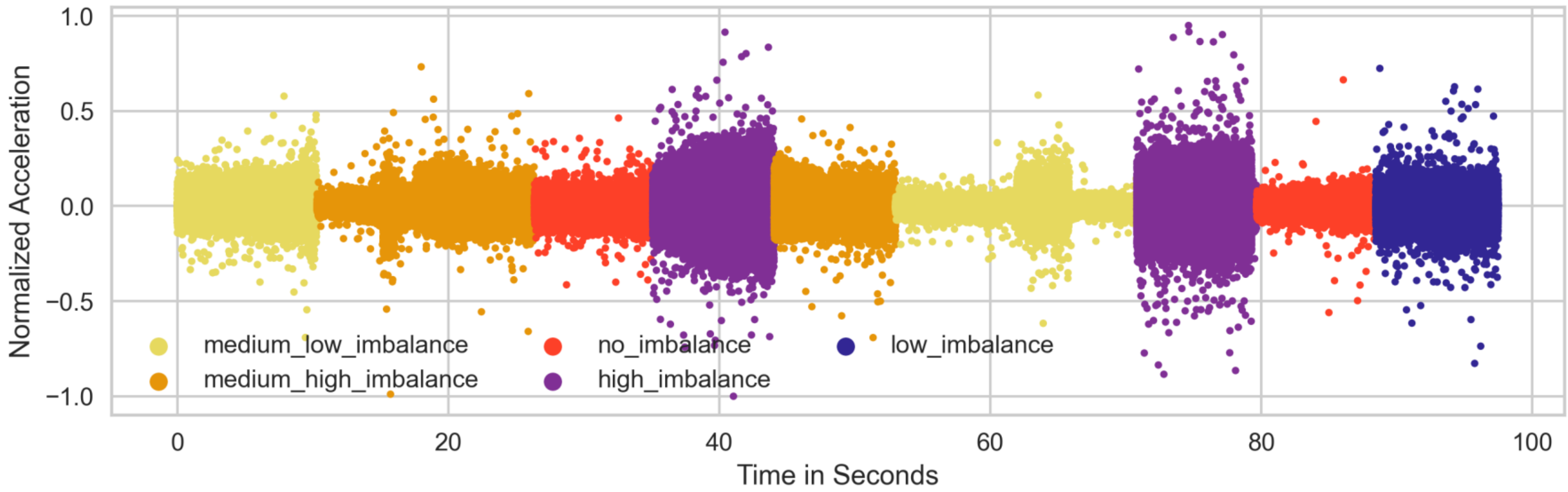
i) Acquired by Siemens for development of anomaly detection and classification algorithms.

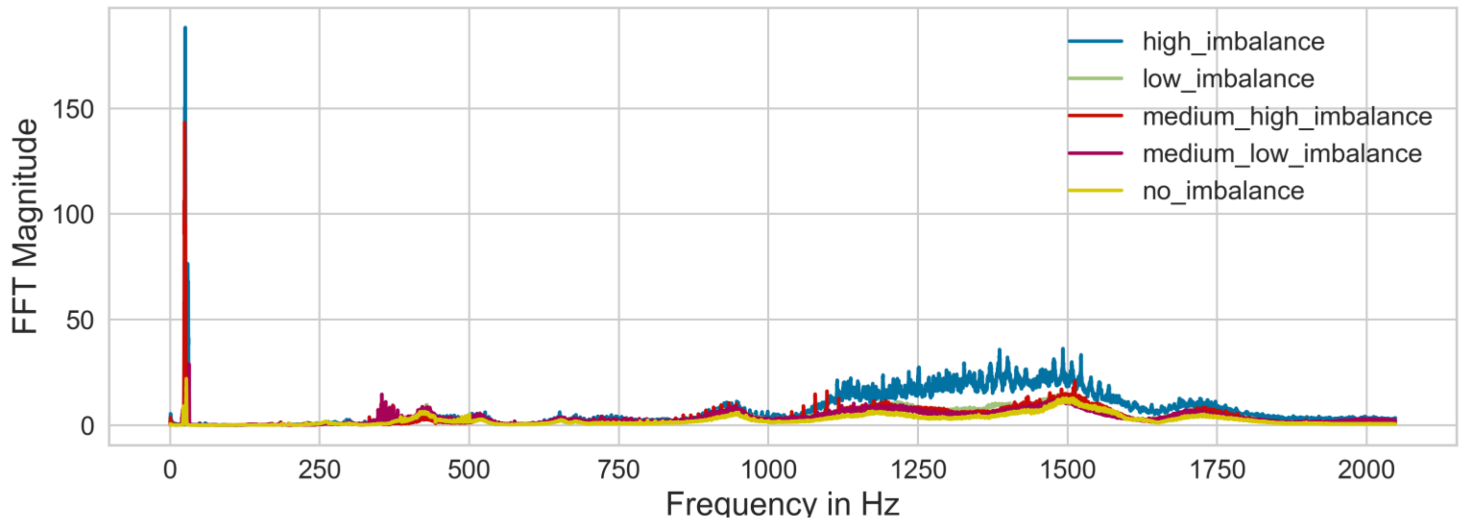
ii) Captured using a test bench with a **centrifugal pump.**

TECHNISCHE UNIVERSITÄT WIEN

SIEMENS

Data Set 2





Data Set 2 in time and frequency domain.

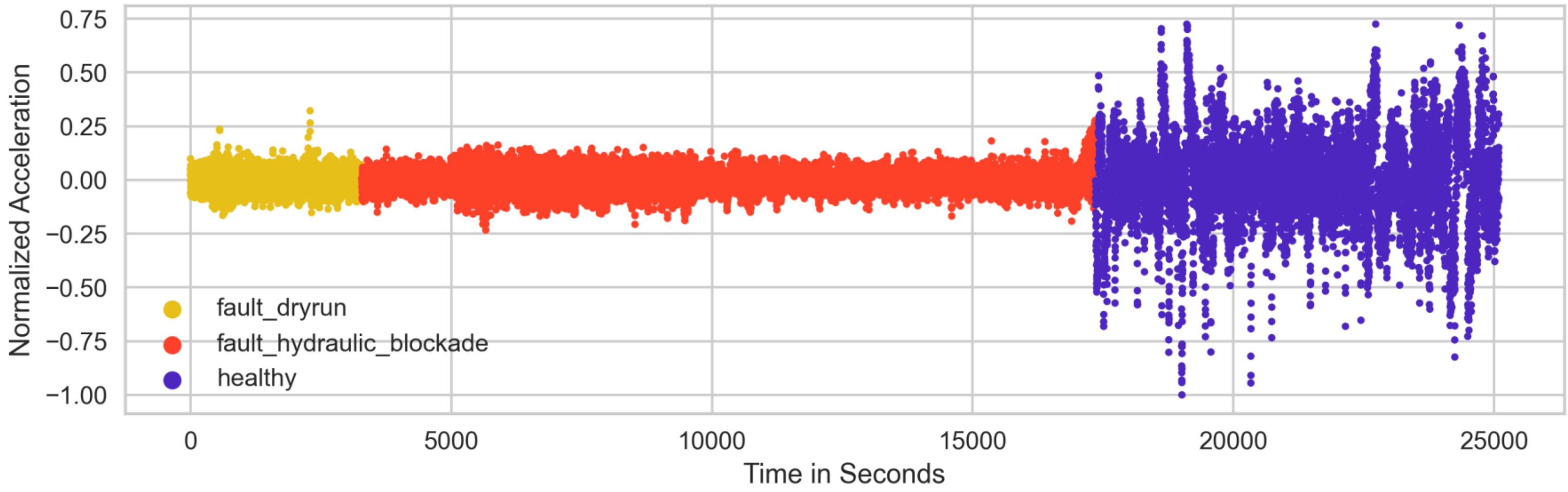
i) Open-source, part of a publication on the development and evaluation of algorithms for imbalance detection

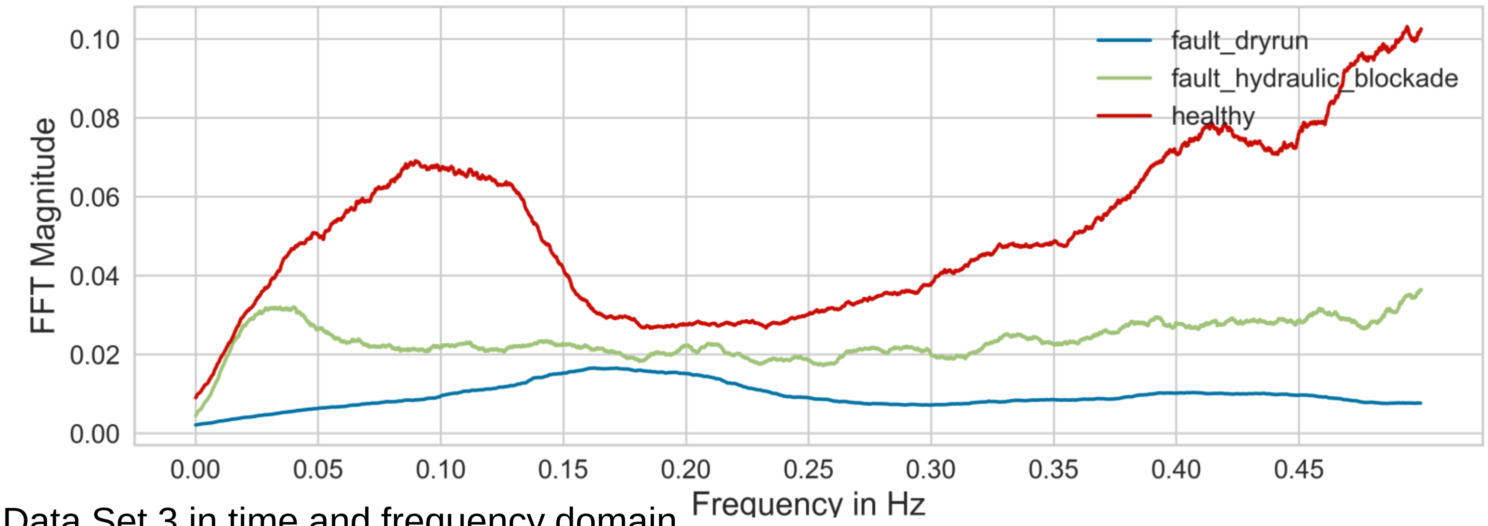
ii) Captured using **imbalanced rotating shafts.**

TECHNISCHE UNIVERSITÄT WIEN

SIEMENS

Data Set 3





Data Set 3 in time and frequency domain.

i) Skoltech Anomaly Benchmark (SKAB), an open-source data set designed for evaluating anomaly detection algorithms.

ii) Captured using a test bench with a **water pump.**

TECHNISCHE UNIVERSITÄT WIEN

SIEMENS

Statistical Feature Extraction

Statistical features were extracted from

i) Time domain (**TD**), derived from **vibrational amplitudes.**

ii) Frequency domain (**FD**), derived from **frequency components.**

Statistical Feature Extraction

Statistical features were extracted from

i) Time domain (**TD**), derived from **vibrational amplitudes.**

ii) Frequency domain (**FD**), derived from **frequency components.**

We used the following features.

i) Arithmetic mean of absolute values (**Abs Mean**)

ii) Median of absolute values (**Abs Median**)

iii) Standard deviation (**Std**)

iv) Interquartile range (**IQR**)

v) Skewness of absolute values (**Abs Skew**)

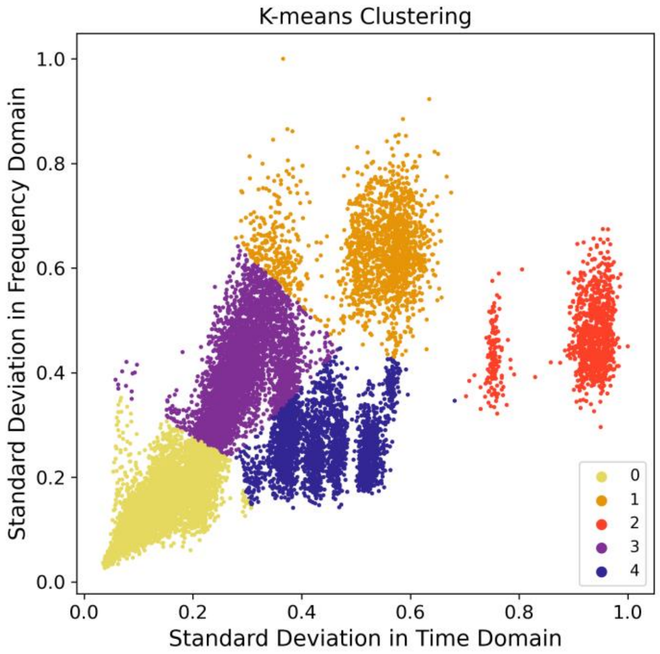
vi) Kurtosis of absolute values (**Abs Kurt**)

TECHNISCHE UNIVERSITÄT WIEN

SIEMENS

Clustering Algorithms

i) **K-means clustering**, which is one of the most popular iterative clustering methods.

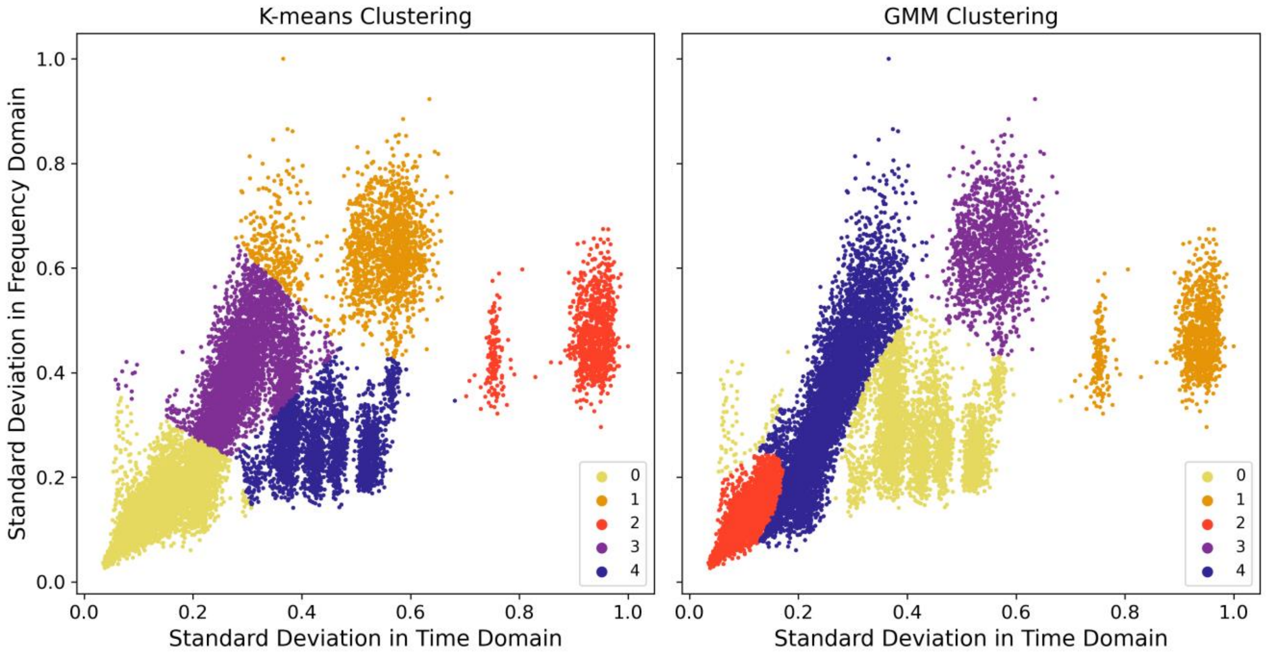


Comparison of different clustering algorithms in the feature space.

Clustering Algorithms

i) **K-means clustering**, which is one of the most popular iterative clustering methods.

ii) **Gaussian mixture model clustering (GMM)**, which models each cluster in terms of a normal distribution.



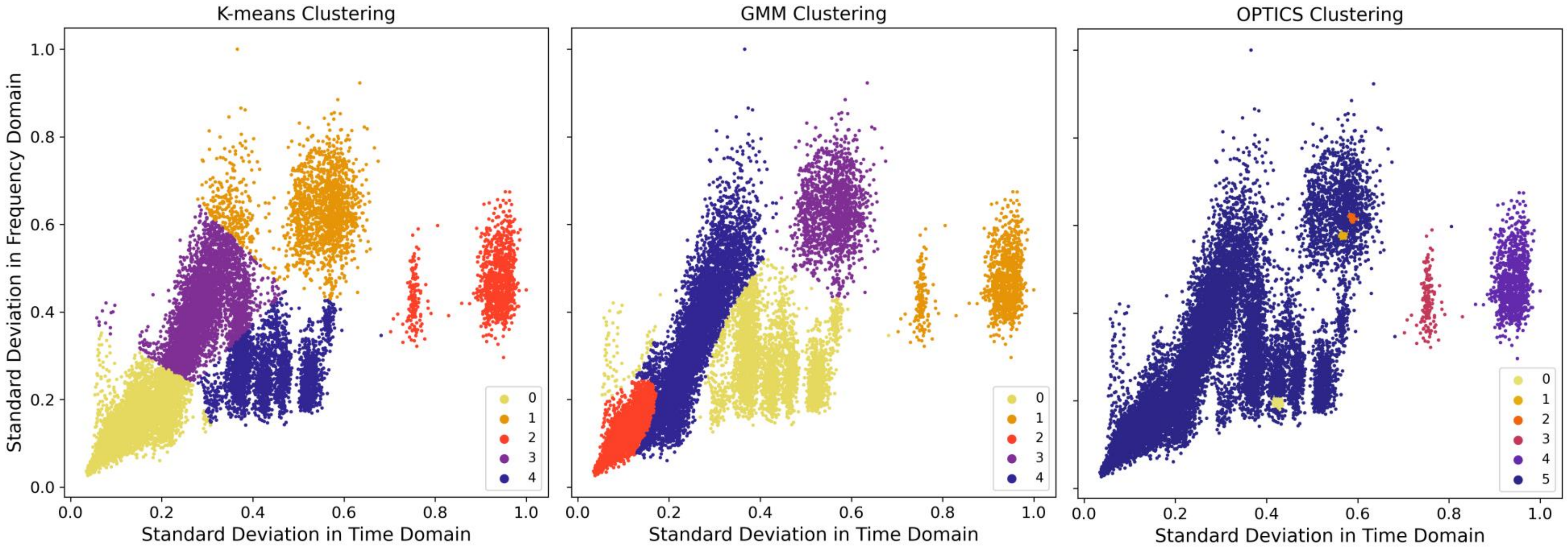
Comparison of different clustering algorithms in the feature space.

Clustering Algorithms

i) **K-means clustering**, which is one of the most popular iterative clustering methods.

ii) **Gaussian mixture model clustering (GMM)**, which models each cluster in terms of a normal distribution.

iii) **Ordering Points To Identify the Clustering Structure (OPTICS)**, which works like an extended **DBSCAN** algorithm for an infinite number of distance parameters.



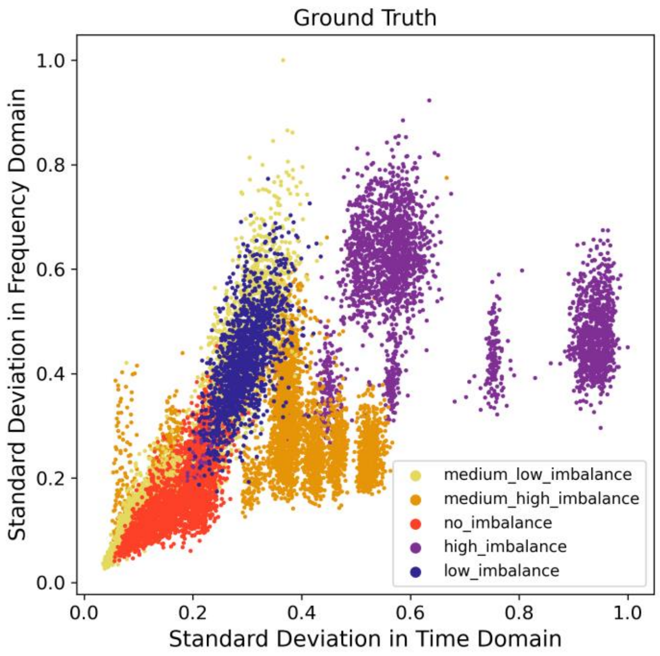
Comparison of different clustering algorithms in the feature space.

Purity

The success of the experiment was measured by the average purity of the resulting clusters.

i) Purity is a measure of the **degree to which clusters only contain a single class.**

ii) It does not penalize an increasing number of clusters.



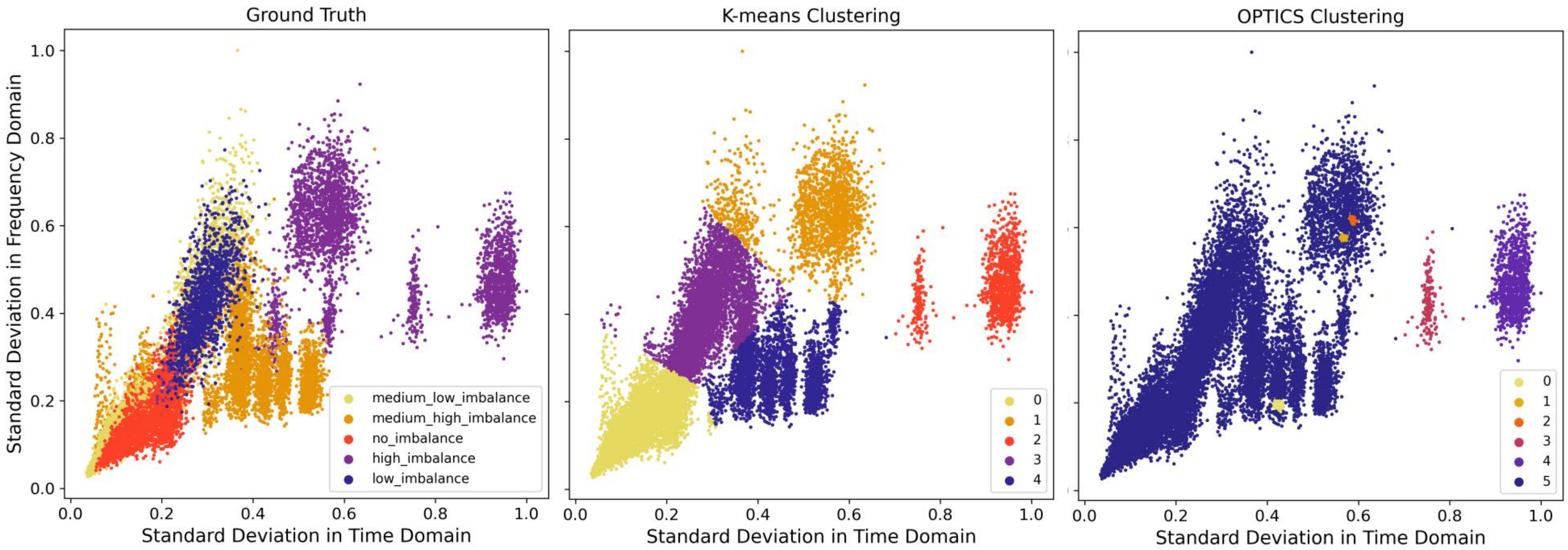
Comparison of ground truth and clusters of different purity in the feature space.

Purity

The success of the experiment was measured by the average purity of the resulting clusters.

i) Purity is a measure of the **degree to which clusters only contain a single class.**

ii) It does not penalize an increasing number of clusters.



Comparison of ground truth and clusters of different purity in the feature space.

TECHNISCHE UNIVERSITÄT WIEN

SIEMENS

Question I

Which **combinations of statistical features and clustering algorithms** perform best for multiple data sets?

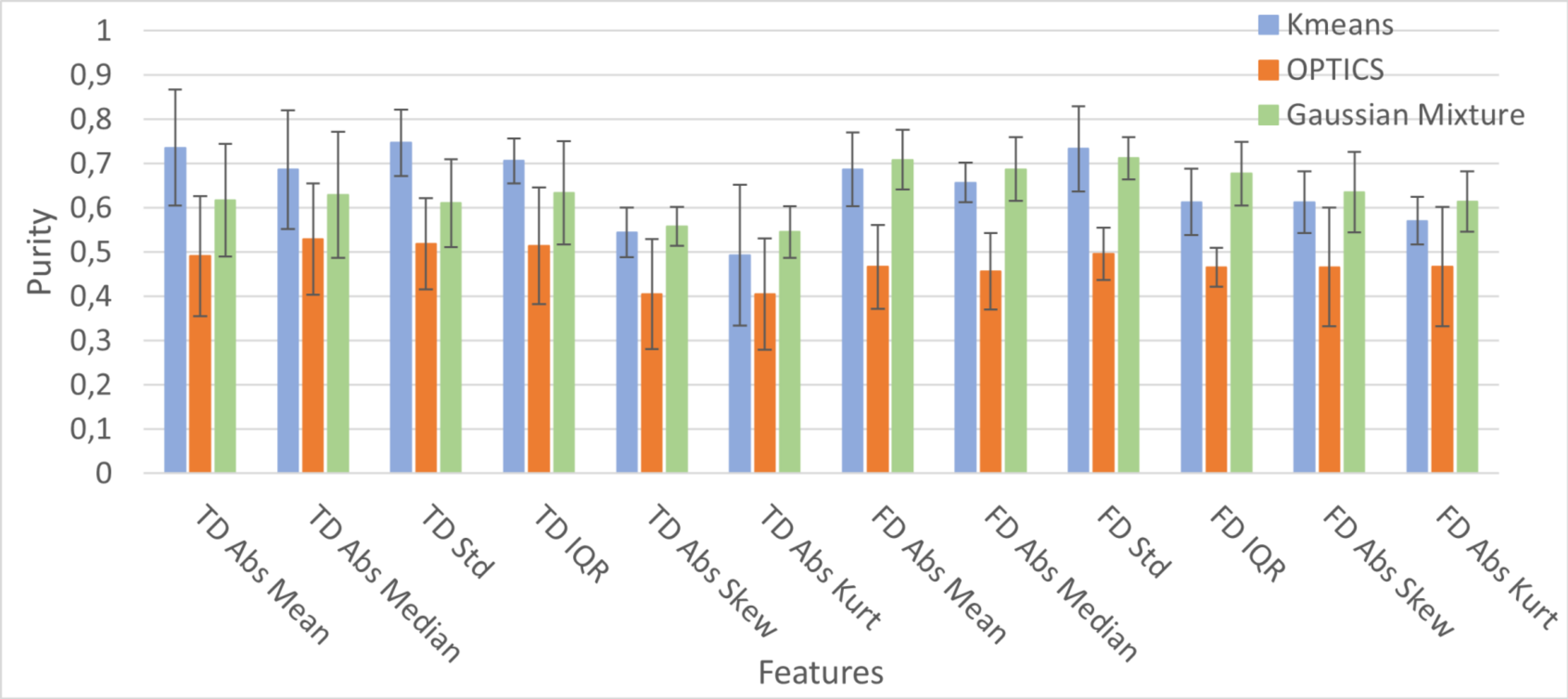
Question I

Which **combinations of statistical features and clustering algorithms** perform best for multiple data sets?

i) K-means performed best.

ii) OPTICS performed worst.

iii) **Lower statistical moments performed better than higher moments.**



Average purity per feature for different algorithms.

Question II

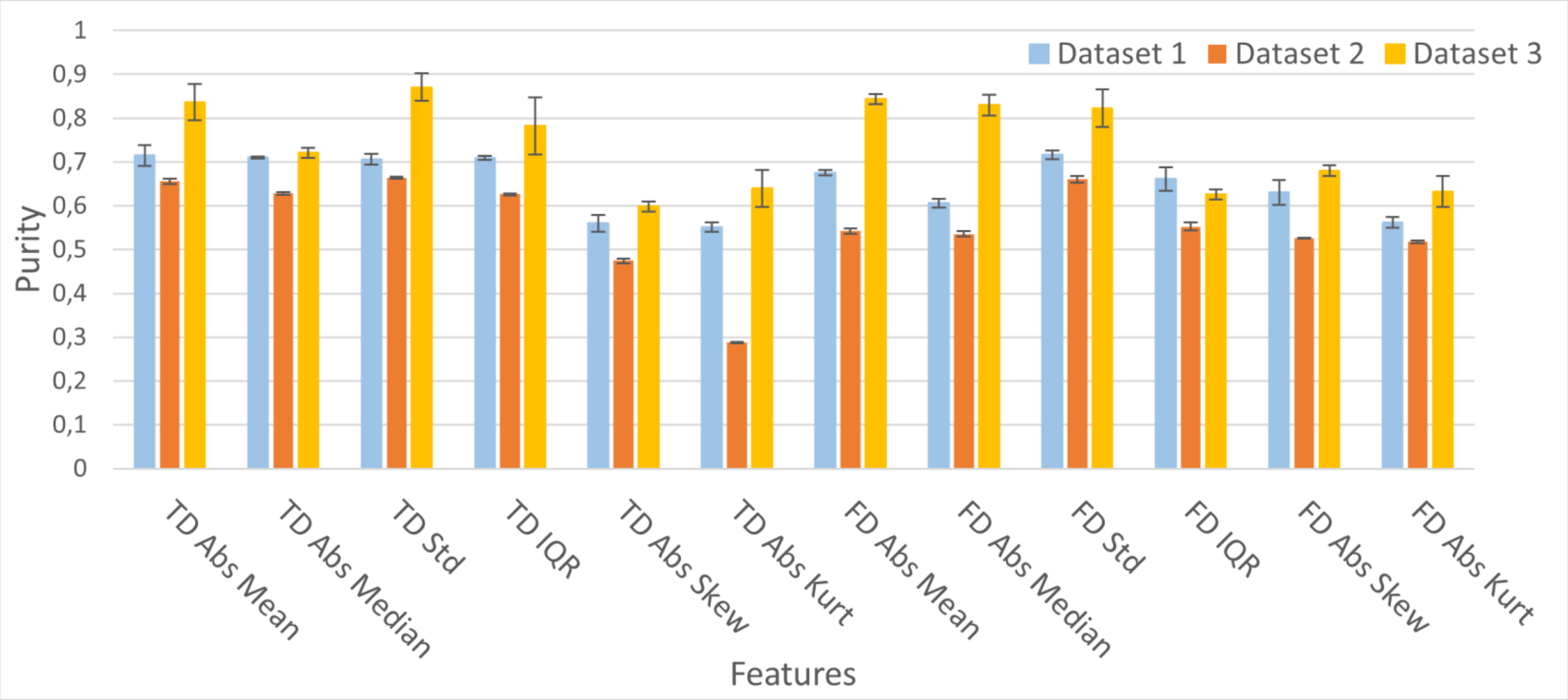
Does the performance of statistical feature and clustering algorithm combinations **generalize for arbitrary data sets**?

Question II

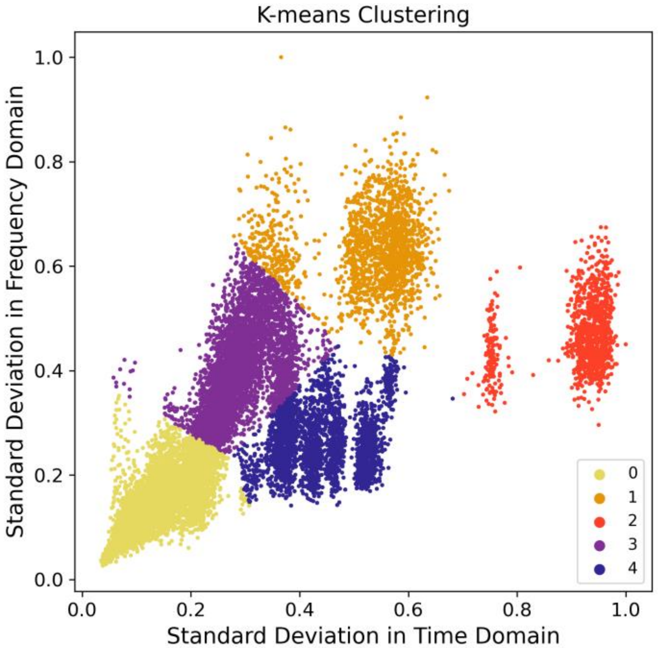
Does the performance of statistical feature and clustering algorithm combinations **generalize for arbitrary data sets**?

i) **Some features appear to be superior in general.**

ii) Does not really generalize for arbitrary data sets.



K-means clustering purity per feature for different data sets.



K-means clustering in the feature space

TECHNISCHE UNIVERSITÄT WIEN

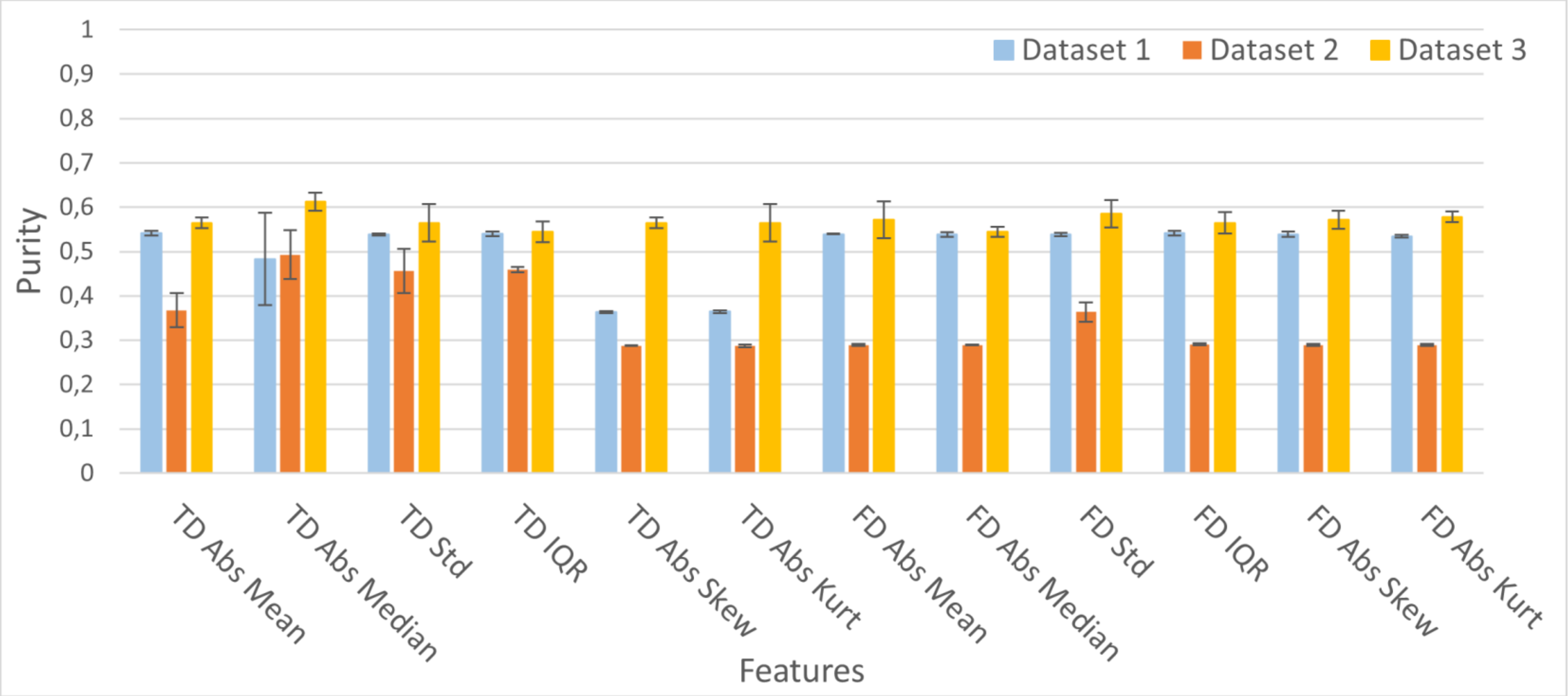
SIEMENS

Question II

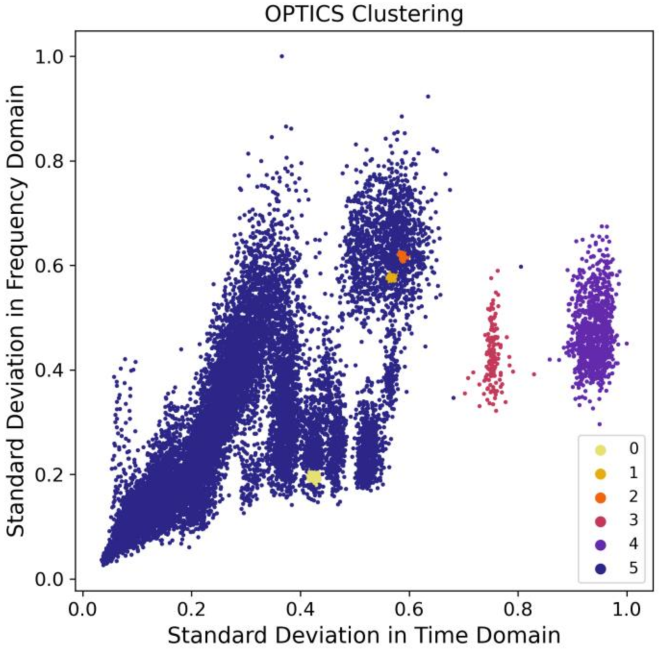
Does the performance of statistical feature and clustering algorithm combinations **generalize for arbitrary data sets**?

i) **OPTICS performed far worse than the other algorithms.**

ii) Could be a result of high variance and low class separability in industrial data.



OPTICS clustering purity per feature for different data sets.



OPTICS clustering in the feature space

TECHNISCHE UNIVERSITÄT WIEN

SIEMENS

Question III

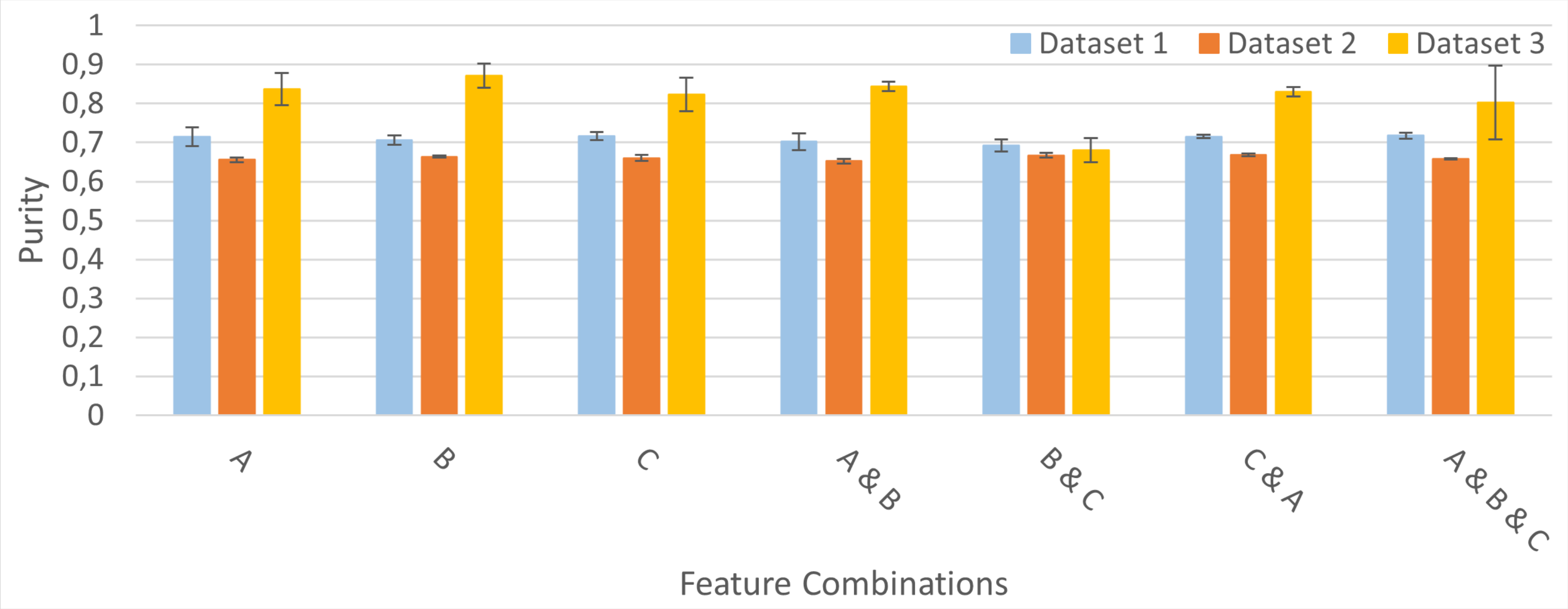
Can the **combination of several different features** improve the performance of the clustering?

Question III

Can the **combination of several different features** improve the performance of the clustering?

i) **It did not.**

ii) Even though they are commonly used in the domain.



K-means clustering purity for feature combinations.

Question IV

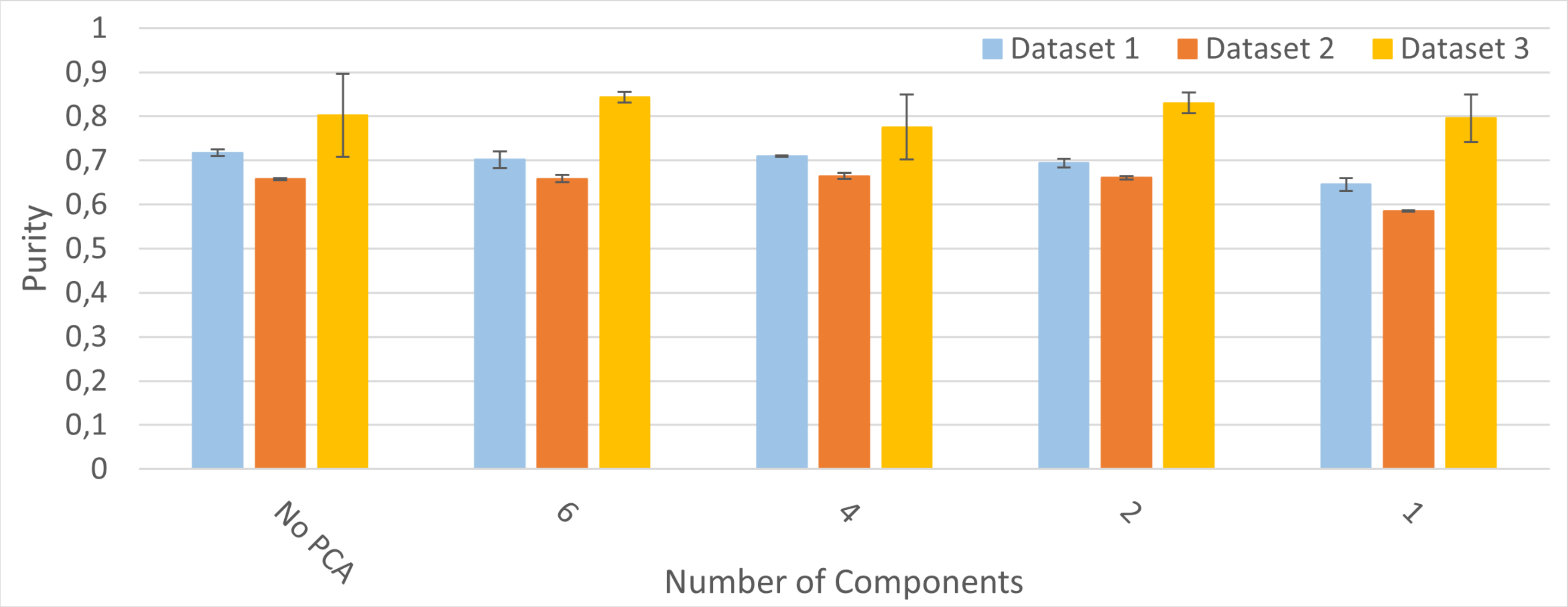
Can principal component analysis (**PCA**) improve the performance of the clustering by selecting the most representative features?

Question IV

Can principal component analysis (**PCA**) improve the performance of the clustering by selecting the most representative features?

i) **It did not.**

ii) It is to note that even just one or two principal components seem to suffice for clustering.



K-means clustering purity for different numbers of principal components.

Question V

How does the **specified number of clusters** affect the performance of the clustering?

Question V

How does the **specified number of clusters** affect the performance of the clustering?

i) Performance of K-means increased for 1.5 times the number of conditions.

ii) **But did not continue to increase with an increasing number of clusters.**

K-means clustering purity for different numbers of principal components.

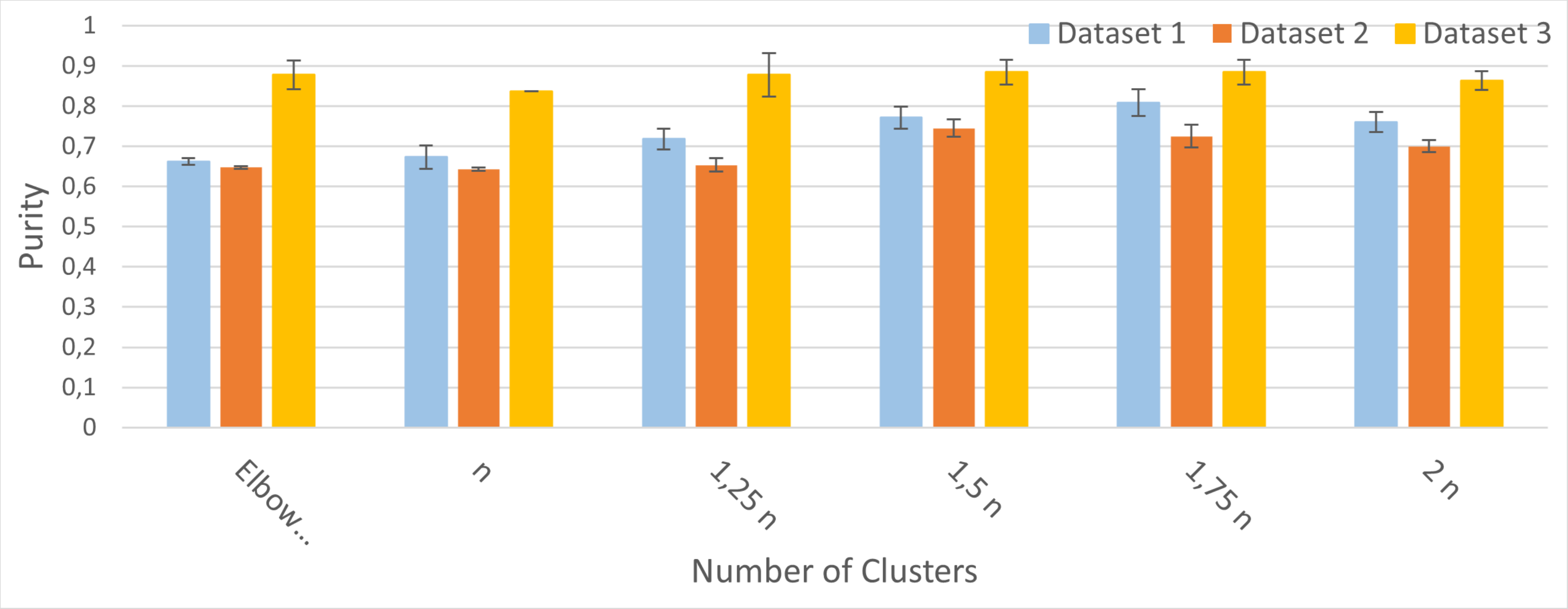
Question V

How does the **specified number of clusters** affect the performance of the clustering?

i) GMM’s performance increased continuously until 2 times the number of conditions

ii) **But declined with an increasing number of clusters.**

iii) Could be a result of GMM not locating any more distinct normal distributions in the data.



GMM purity for different numbers of principal components.

TECHNISCHE UNIVERSITÄT WIEN

SIEMENS

Conclusion

What did we learn?

i) In vibration data, **lower statistical moments are more important.**

ii) K-means and GMM perform far better than OPTICS for this data.

iii) Limited improvements from feature combinations and PCA.

iv) **Ideal number of clusters of about 1.5 to 1.75** times the number of conditions.

Conclusion

What did we learn?

i) In vibration data, **lower statistical moments are more important.**

ii) K-means and GMM perform far better than OPTICS for this data.

iii) Limited improvements from feature combinations and PCA.

iv) **Ideal number of clusters of about 1.5 to 1.75** times the number of conditions.

What are the limitations?

i) **Only three different data sets** were used.

ii) **Only three select clustering algorithms** were used.

iii) Only three tests per experimental setting were run.

TECHNISCHE UNIVERSITÄT WIEN

SIEMENS

Conclusion

What did we learn?

i) In vibration data, **lower statistical moments are more important.**

ii) K-means and GMM perform far better than OPTICS for this data.

iii) Limited improvements from feature combinations and PCA.

iv) **Ideal number of clusters of about 1.5 to 1.75** times the number of conditions.

What are the limitations?

i) **Only three different data sets** were used.

ii) **Only three select clustering algorithms** were used.

iii) Only three tests per experimental setting were run.

Future work.

i) **Increasing the number of data sets** for better conclusions about generalizability.

ii) **Increasing the number of clustering algorithms**.

TECHNISCHE UNIVERSITÄT WIEN

SIEMENS

Thank you for your attention!

Comparison of Clustering Algorithms for Statistical Features of Vibration Data Sets

Philipp Sepin, Jana Kemnitz, Safoura Rezapour Lakani, Daniel Schall

Philipp Sepin
Distributed-Artificial Intelligence-Systems Research Group, Siemens Technology
Vienna University of Technology