Naval Information Warfare Center PACIFIC

TECHNICAL REPORT 3323
NOVEMBER 2023

An Investigation of Time Series Embeddings for Topological Data Analysis

Dean Lee
Jamal Rorie, Ph.D.
Andrew Sabater, Ph.D.
NIWC Pacific

Naval Information Warfare Center (NIWC) Pacific
San Diego, CA 92152-5001

This page is intentionally blank.

TECHNICAL REPORT 3323
NOVEMBER 2023

An Investigation of Time Series Embeddings for Topological Data Analysis

Dean Lee
Jamal Rorie, Ph.D.
Andrew Sabater, Ph.D.
NIWC Pacific

Administrative Notes:
This report was approved through the Release of Scientific and Technical Information (RSTI) process in September 2023 and formally published in the Defense Technical Information Center (DTIC) in November 2023.

NIWC Pacific
San Diego, CA 92152-5001

NIWC Pacific
San Diego, California 92152-5001

P.M. McKenna, CAPT, USN
Commanding Officer

M.J. McMillan
Executive Director

ADMINISTRATIVE INFORMATION

The work described in this report was performed by the Data Science, Engineering, and Analytics Branch of the C2 Science, Technology, and Engineering Division, Naval Information Warfare Center (NIWC) Pacific, San Diego, CA. The In-House Laboratory Independent Research (ILIR) program provided funding for this work.

Released by
Emily Nguyen, Division Head
C2 Science, Technology, and Engineering Division

Under authority of
Scott Crellin, Department Head
Command & Control and Enterprise Engineering Derpartment

Editor: MRM

EXECUTIVE SUMMARY

Topological data analysis (TDA) is an important part of a data scientist's tool box; through the extraction of topological information, TDA provides an automated means of feature engineering. Thus, TDA may be an appropriate tool for fault analysis of mechanical systems, where the methodologies have traditionally relied on expert intuition and signal processing techniques. However, appropriate embedding of the data must be defined before TDA can be applied. In this paper, we investigate several different embeddings of time series data for the application of TDA for fault analysis, and show that features engineered by TDA improve fault classification accuracy.

This page is intentionally blank.

CONTENTS

Figures

Tables

1. INTRODUCTION

Modern mechanical systems are instrumented with sensors that provide information for system health diagnostics. Such information could be used to infer health state of hard-to-reach components, so that appropriate maintenance actions can be performed to minimize maintenance cost and maximize system uptime. Traditionally, algorithms developed for these systems have relied on expert intuition and signal processing techniques [1–4] to develop fault features for diagnosis. More recently, with the advent of deep learning, the industry has seen the development of sophisticated machine learning algorithms for fault analysis [5–9]. Deep learning techniques, however, do not usually provide built-in explainability, and may require other means to provide intuition for the results.

Topological data analysis (TDA) provides a new set of tools for automated feature extraction that often provide better interpretation of the results. Persistent homology [10] in particular, is a technique used to transform the latent topological information resident in the data into features. Lee et al. [11] demonstrated the feasibility of using TDA to determine faults in roller element bearings, and showed that the visualization of the topological features of vibration signals may convey important information about the health of the component. The prerequisite step of TDA application, however, is the selection of an appropriate embedding. In particular, the visibility graph [12] was used in Reference [11] to embed raw vibration data from which point clouds are formed; the topological features extracted from the point clouds are then used to train machine learning models to classify faults. While the approach created models with high accuracy, it leaves open the question whether there are other embedding techniques which may be more appropriate.

In this paper, we explore different embedding techniques with which to transform time series data for topological data analysis. The raw vibration data from a rolling element bearing data set are embedded into a metric space, from which topological features are extracted. Machine learning models are then built with the features extracted from each embedding to classify faults. Additionally, the models are also evaluated against artificially downsampled data, which are common in certain domains where data storage capacity is limited, to determine the feasibility of TDA in an degraded operational environment. Finally, the correlation of the topological features to vibration metrics in the time domain are investigated to determine the connection between the topological features extracted from the various embeddings to the classical signals processing methods. These correlations show that TDA is automating the creation of features that are similar to the vibration metrics.

This page is intentionally blank.

2. MOTIVATION

In [11], a visibility graph-based topological data analysis method was proposed, and applied on the bearing vibration data from the Society for Machinery Failure Prevention Technology (MFPT) and the Center for Intelligent Maintenance Systems (IMS) to extract topological information for machine learning. It was shown that a machine learning algorithm can be built to detect various bearing faults, and that distinct topological structures were generated from fault-induced vibrations.

However, visibility graph [12] is just one method to embed time series data, and the topological information extracted from point clouds are specific to the embedding. Thus, an investigation of other embedding techniques and the topological features that they generate is warranted in order to better understand the efficacy of the proposed method. Moreover, other embeddings may produce structures which provide deeper insight for fault diagnosis and potentially show connections to various vibration metrics. This is crucial in establishing the proposed method as a legitimate tool for vibration analysis.

3. PRELIMINARIES

3.1 PERSISTENT HOMOLOGY

Persistent homology [13] is a method for characterizing the shape of data. Intuitively, the points in a point cloud are balls with fixed radii; a simplex is constructed by forming edges where the surface of the balls first intersect. By increasing the radii, or the filtration parameter, at a uniform rate, simplicial complexes are formed and absorbed. In this paper, the Vietoris-Rips complex [10] is used as an approximation for the formation of simplicial complexes, due to its computational efficiency. Specifically, Vietoris-Rips complexes are formed by considering pairwise intersections of the balls in a point cloud; the details of the construction of Vietoris-Rips complexes can be found in [14].

The number of structures formed during the filtration process are captured in homology classes, H_i, $i \geq 0$. Intuitively, H_0 represents the number of connected components, H_1 represents the number of loops, etc., in the simplicial complex at a given filtration value. More background information about homology classes can be found in [15].

A persistence diagram provides a summary of the filtration process by capturing the formation and destruction of simplicial complexes throughout the process. More precisely, a persistence diagram is the multiset $\{(b_i, d_i)\}_{i \in I}$, where (b_i, d_i) denote the birth and death time of simplicial complex i. Point summaries can be extracted from persistence diagrams [16–18] and used as features for machine learning algorithms. In this paper, we use the weighted persistence silhouettes [16] generated from persistence diagrams. In particular, let $\{(b_i, d_i)\}_{i \in I}$ be the birth/death pairs of a persistence diagram, then the silhouette of the persistence diagram is defined by

$$\phi(t) = \frac{\sum_{i \in I} |d_i - b_i| \Lambda_i(t)}{\sum_{i \in I} |d_i - b_i|},$$

where $\Lambda_i(t) = \max\{0, \min\{t - b_i, d_i - t\}\}$. Intuitively, the silhouette diagram helps to show where long-persisting structures are concentrated, which is a function of the underlying data. Other point summaries that are used as machine learning features in this paper are the persistence landscape [16], persistence entropy [19], and the Betti curve [14]. We use the implementation of `giotto-tda` [20] to compute these point summaries.

3.2 EMBEDDINGS

Time series data must be embedded in a space in which point clouds could be formed, and from which topological data analysis can be performed. Crucially, an appropriate distance metric must also be determined for the space. In general, a sliding window of values is created across the time series data; the values of the window are embedded using various methods. The embedding methods investigated in this analysis are listed below.

3.2.1 SIMPLE EMBEDDING

In this embedding, each window in the time series is treated as a vector. It is customary to use the Euclidean distance as the default metric to quantify the difference between vectors, and indeed, the Euclidean distance is the default of many machine learning software packages. However, it is not always the appropriate distance measure. It is noted in [21] that the Euclidean distance breaks down even in relatively low dimensions, and it is recommended that the Manhattan distance be used instead as the default measure. We will thus use the Manhattan distance as the measure for the vector space embedding.

3.2.2 PROBABILITY DISTRIBUTIONS

The values of each window are normalized and binned to create a vector of frequencies, or empirical distributions of the data. We use the Jensen-Shannon divergence [22] metric to quantify the similarity between two distributions.

3.2.3 VISIBILITY GRAPHS

The visibility graph [12] provides a framework for embedding time series data into graphs. Intuitively, the time series data is a series of peaks and valleys. Every node in the graph corresponds to a sample from the time series, and an edge is constructed between two nodes if their corresponding peaks in the time series are visible to each other. The visibility graphs may quantify local and global information in the form of graph structures: a highly connected community of nodes in the graph indicate that the corresponding values in the time series are close in value (hence, visible to each other); the links that connect the disparate communities are defined by the spikes in the time series signal, which "wall" in the communities. In this sense, the visibility graphs capture interesting information from the time series signal. Topological data analysis uses the graphs as a metric space to extract topological features that quantify these information.

In particular, given a time series $\{(t_i, y_i)\}$, a vertex exists for every corresponding t_i, and an edge exists between t_a and t_b if for any $(t_c, y_c), t_a < t_c < t_b$,

$$y_c < y_b + (y_a - y_b)\frac{t_b - t_c}{t_b - t_a}.$$

There are different ways to calculate the distance between two nodes, and the right choice for the metric has a big performance impact on the models. For this analysis, undirected graphs are generated, it was discovered that the Euclidean distance between two points perform the best, and is used as the edge weights for the visibility graphs. The visibility graph implementation of `ts2vg` [23] is used in the analysis.

3.2.4 TAKENS EMBEDDING

Takens embedding [24] embeds the time series data into higher dimensions. More formally, given a time series $f(t)$, an embedding dimesion d, a time delay τ, and time t_i, the embedding produces vectors of the form

$$(f(t_i), f(t_i + \tau), f(t_i + 2\tau), ..., f(t_i + (d-1)\tau)).$$

These vectors form points clouds, and from which persistent homology is applied to extract topological information. We use the `giotto-tda` [20] implementation of Takens embedding in this paper.

3.2.5 SYMBOLIC AGGREGATION APPROXIMATION

The Symbolic Aggregation approXimation (SAX) algorithm is a time series dimensionality reduction method introduced in [25]. The method transforms a time series into a series of symbols, from which natural language processing techniques can be applied for feature extraction.

SAX starts by standardizing the time series data. Sliding windows of size w are created on the time series data, and within each window, m number of equal-sized segments are created; the mean value of each segment is computed; the m parameter determines the word size of the SAX algorithm. An alphabet of size n is also chosen; the values of n corresponds to the number of equal-sized bins under the $N(0, 1)$ distribution, and each bin is assigned a letter. Finally, the mean values computed for each of the m

segments are converted to letters according to where the values fall in the bins defined for $N(0, 1)$, and the SAX words are formed in this manner.

In this paper, we built upon the previous work by expanding the use of SAX an embedding algorithm for time series data. For each of the windows of data, a series of words are formed, and the frequency vector of the words are computed and used to form point clouds. Persistent homology is then applied on the point clouds for feature extraction.

4. METHODOLOGY

The notional pipeline for generating the topological features is shown in Figure 1.

A sliding window is created on the raw vibration data. These windows are embedded into spaces proposed in Section 3.2, and from which point clouds are constructed using the appropriate distance measures. In the topological feature extraction step, persistent homology is computed on the point clouds, from which persistence diagrams and point summaries are generated as features for machine learning models. We use the persistent homology implementation of giotto-tda [20] for all analyses, and we use the `scikit-learn` implementation of Random Forest with default settings as the base machine learning model. Finally, the data is split into train, test, and validation sets, and the validation set is used to find optimal parameters for each of the proposed embeddings.

Figure 1. Notional topological data analysis pipeline.

5. DATA

We analyze the Case Western Reserve University (CWRU) bearing data set [26] to examine the effects of the proposed embeddings on the generated topological features. The data set is made up of vibration measurements from healthy bearings and bearings with seeded faults. The seeded faults are introduced via the electro-discharging machining methods at various depths. Three types of faults are introduced: inner raceway fault, ball-bearing fault, and outer raceway faults. The vibration measurements are collected at the drive-end and the fan-end of the test apparatus, and the data are collected at various motor loads. While the data set contains drive end and fan end bearing fault information, we only consider the drive end faults for the following analyses as the data is more complete. Furthermore, only the data collected at 12,000 samples/second are considered to ensure uniformity of analysis, as the baseline data is only collected at 12,000 samples/second.

The main objective of the analysis is to determine the effectiveness of topological data analysis for fault classification, and three sets of experiments are conducted using the CWRU data set. In the first set of experiments, models are built from topological features extracted from the various embeddings with which to classify different faults at various motor loads and fault depths.

In the second set of experiments, artificial downsampling is applied to the data set to simulate real-world scenarios. In particular, data may be stored in decimated form in order to reduce storage footprint in certain operational environments. Importantly, a downsampling of factor n used here is meant that every n-th sample is kept from the original data. In these experiments, we investigate the performance of the models at various levels of downsampling factors to understand how topological data analysis may perform in real-world settings.

Finally, the connection to some classical vibration metrics are considered. In the third set of experiments, common vibration metrics, such as peak acceleration, root-mean-square, crest factor, and kurtosis are correlated with the topological features to show possible connections to these metrics. In particular, these may shed light on how TDA may be automating the generation of important vibration metrics.

6. RESULTS

6.1 FULL DATA RESULTS

The model performance based on the topological features extracted from the different embeddings are shown in Figure 2. In general, the models perform better at 7 and 21 millimeter of fault depths. It can be seen that the distribution embedding performs the best out of the five embeddings considered: the model performance is relatively stable and accurate across different loads and fault depths. This suggests that by binning the data points appropriately, the differences in the vibrations among the various fault categories can be quantified as distinct topological features for machine learning.

The simple embedding model performs relatively well at fault depth of 7 millimeters; at other fault depths the performance starts to degrade. This may be due to that vectors aren't capturing enough of the intrinsic differences among the various faults.

The visibility graph model also has inconsistent performance, which may be attributed to its sensitivity to small changes in the vibration measurement; that is, entirely new communities in the graph may appear or disappear due to slight changes in the vibration measurement. Furthermore, it stands to reason that if the vibration sensors aren't calibrated correctly, the visibility graph method may yield poor performance.

The SAX embedding model appears to be relatively unstable, which may be due to the standardization of the data for forming the words, and that may mask the subtle differences in the changes in the vibrations across the various fault depths.

The Takens embedding model has the worst performance. This suggests that this embedding technique does not generate features which generalize well across different fault categories.

6.2 DOWNSAMPLED DATA RESULTS

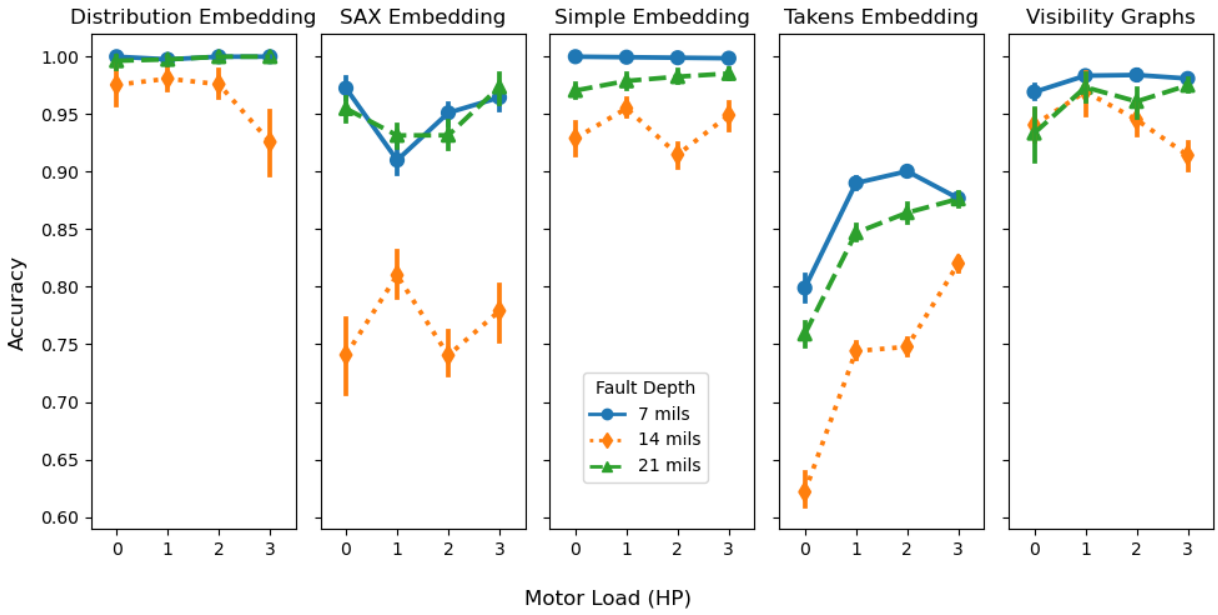
The model performance at various levels of downsampling of the data is shown in Figure 3 and Figure 4 for motor loads 0 and 3 HP, respectively, to highlight changes in performance at extreme ends of motor load values; results for the other motor loads are not shown for brevity.

In general, we see that the simple embedding model performance stays relatively stable over various downsampling factors, as well as across different loads. The distribution embedding model, however, becomes more and more unstable as the data become sparser. One reason may be that as the data become sparser, the empirical distributions no longer represent the underlying distribution.

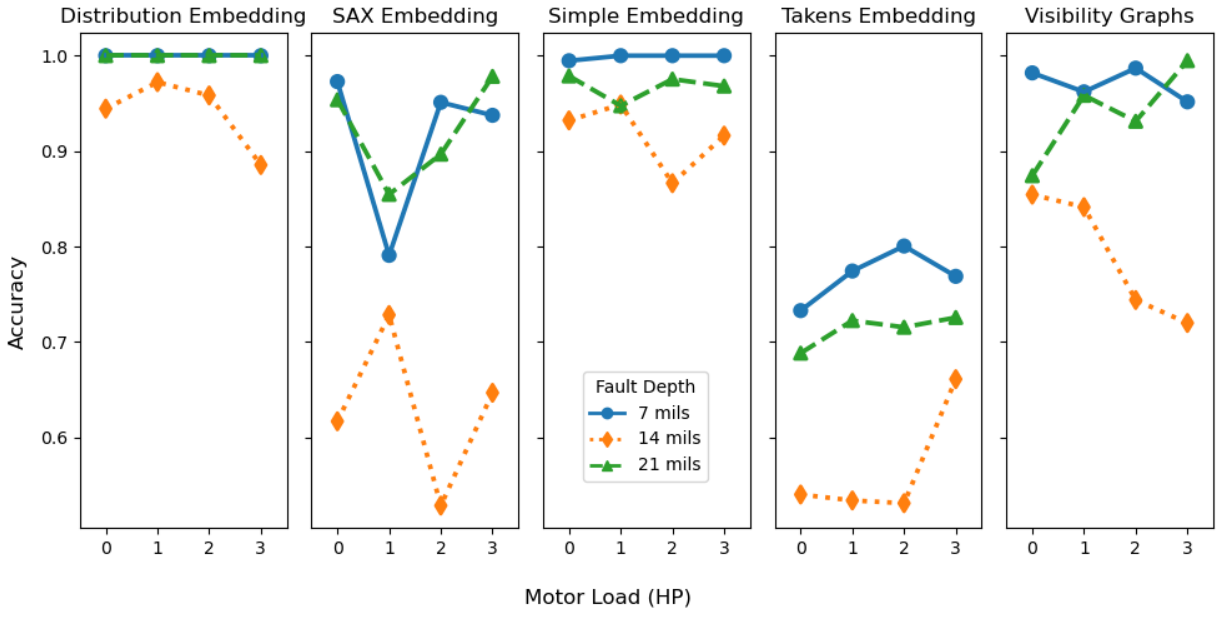
The SAX embedding model also suffers from the same issue that causes the instability in the distribution embedding model, albeit the problem is worse. As the data becomes sparser, the standardization of the sparse data obscures the changes in vibrations, and SAX representations are poor features for distinguishing among various faults.

The visibility graph model also has spots of instability, as the downsampling factors vary. In particular, there is a dip in performance at downsampling factor of 3, after which the performance improves until downsampling factor of 7. Surprisingly, the visibility graph model performance, even at extreme sparsity, is on par with the full data set. This suggests that the graph may be encoding intrinsic properties of the vibration signal.

The Takens embedding model again fare the worst performance, although it was already shown in the previous section that this particular embedding technique may not be suitable for generating topological features that distinguish among various faults. On the other hand, the model performance appears to be more stable than that of the SAX embedding model over various downsampling factors.



(a) Model cross-validation performance.



(b) Model test performance.

Figure 2. Model performance on full data.

6.3 CORRELATION TO VIBRATION ANALYSIS METRICS

Given a window of vibration signals $x = (x_0, x_1, ..., x_{n-1})$, some vibration metrics which are frequently employed for analyses in the time domain are the following.

1. **Peak acceleration**: $\max |x_i|$.
2. **Root mean square (RMS)**: $\sqrt{\frac{1}{n}\sum_i x_i^2}$.
3. **Crest factor**: $\frac{\max |x_i|}{\sqrt{\frac{1}{n}\sum_i x_i^2}}$.
4. **Kurtosis**: $E\left[(\frac{x-\mu}{\sigma})^4\right]$.

In general, these metrics are different ways of quantifying spikes in the signal. When coupled with expert intuition, these values could be used to diagnose underlying conditions.

We examine the correlations between these classical vibration metrics and the topological features extracted from the embeddings. In Figure 5, it can be seen that the homology group 0 of the persistence silhouette generated from the visibility graph and the Takens embeddings have high correlations to the peak acceleration and RMS metrics.

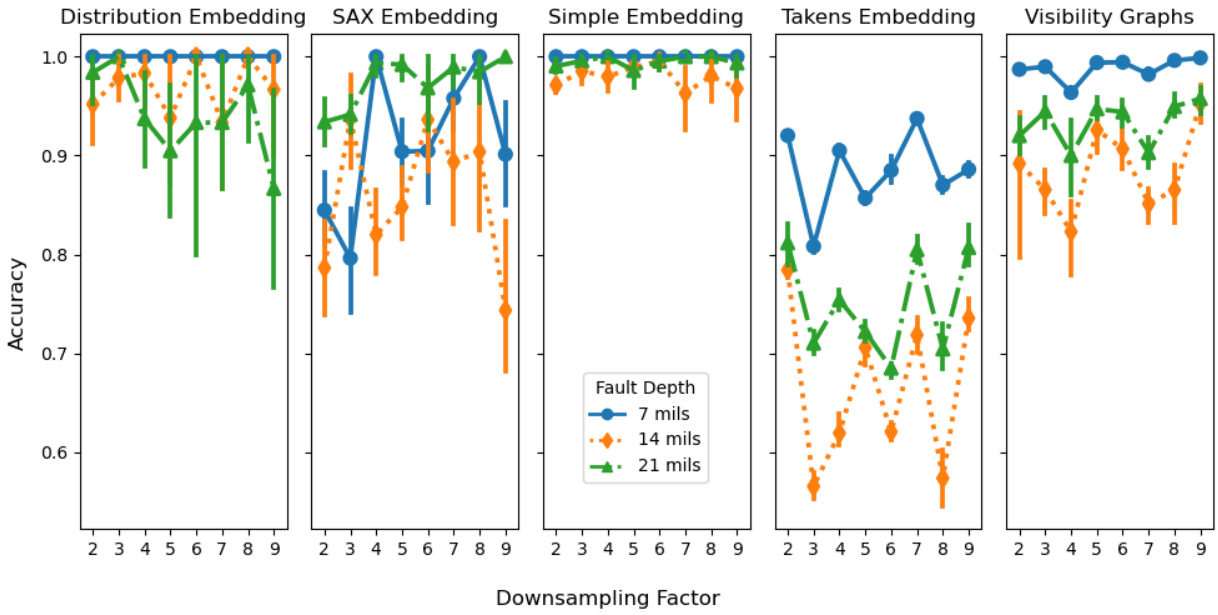
We examine the correlations in a systematic manner by setting a threshold of 0.8 on the absolute value of the correlation coefficients, and vary the sliding window size from 50 to 2,000. On one extreme, Table 1 shows that only the visibility graph and the Takens embedding have generated topological features which correlate to some vibration metrics, which are RMS and peak accelerations. On the other hand, at window size 2,000 (results not shown), most of the topological features from all embeddings correlate to the vibration metrics under consideration.

In Table 2, we show the results for window size 1,000, which is the point at which all of the embeddings start generating topological features which correlate to some vibration metrics. In particular, the visibility graph and distribution embedding generate features which correlate to all four vibration metrics. We also note that the topological features of homology group 1 also are correlated to vibration metrics. Moreover, most of the correlations did not develop until window size of 1,000 was used, and we note that persistence entropy is never correlated with any vibration metrics.

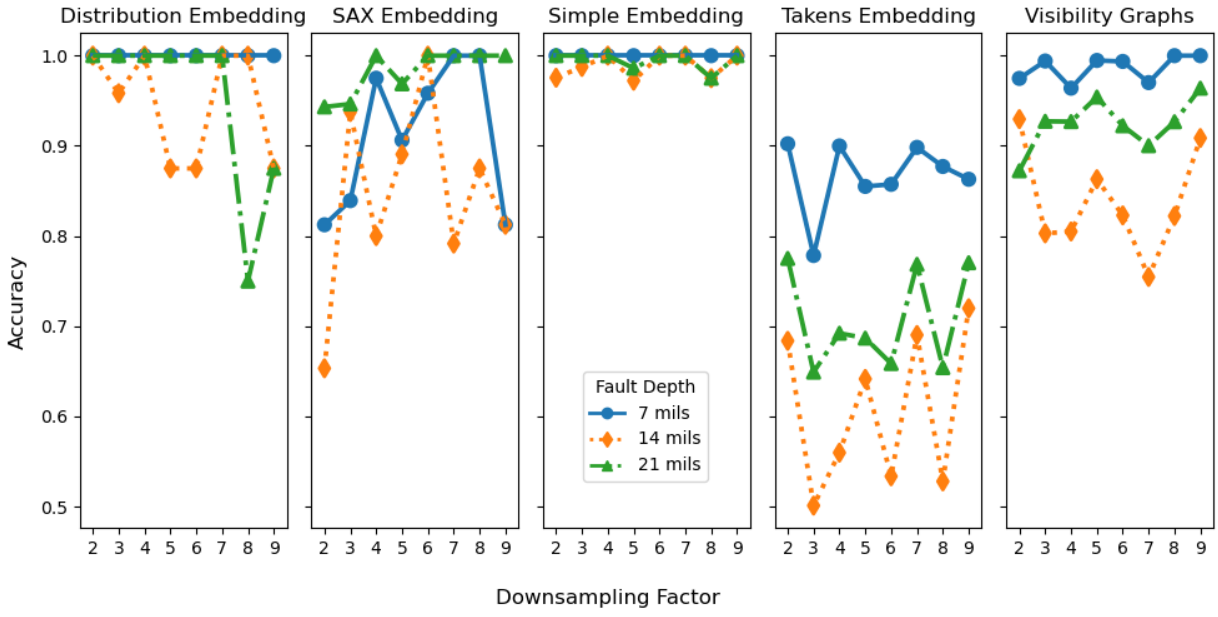
Embedding	Topological Features	Homology Group	Vibration Metric
Simple Embedding			
SAX Embedding			
Visibility Graph	Silhouette, Landscape, Betti	0	RMS,Peak Acc.
Distribution Embedding			
Takens Embedding	Silhouette, Landscape, Betti	0	RMS, Peak Acc.

Table 1. Correlated topological features and vibration metrics at window size 50. The blank entries indicate no strong correlations were detected between the topological features and vibration metrics. The values in the Homology Group column indicates whether the correlated topological features come from homology group 0 or homology group 1.

Closer examinations of the correlations reveal that the correlations occur predominantly in the ball-bearing fault data set, with the exception of the visibility graph and Takens embedding, which generate topological features that correlate with vibration metrics across most of the fault conditions, loads, and depths, even at small window sizes.

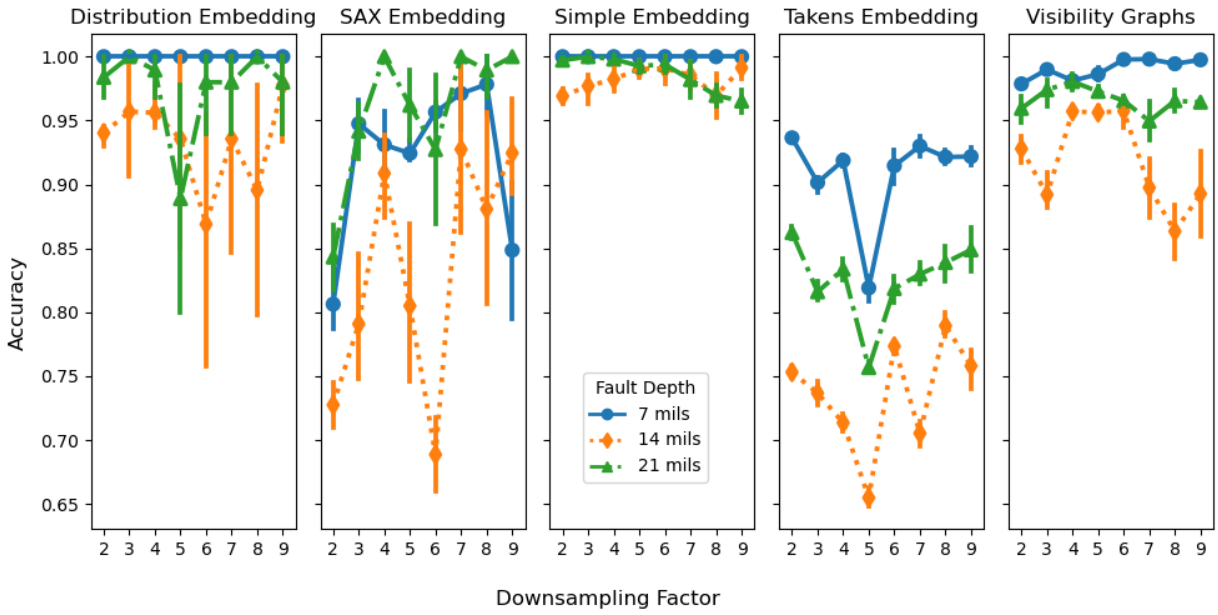


(a) Model cross-validation performance.

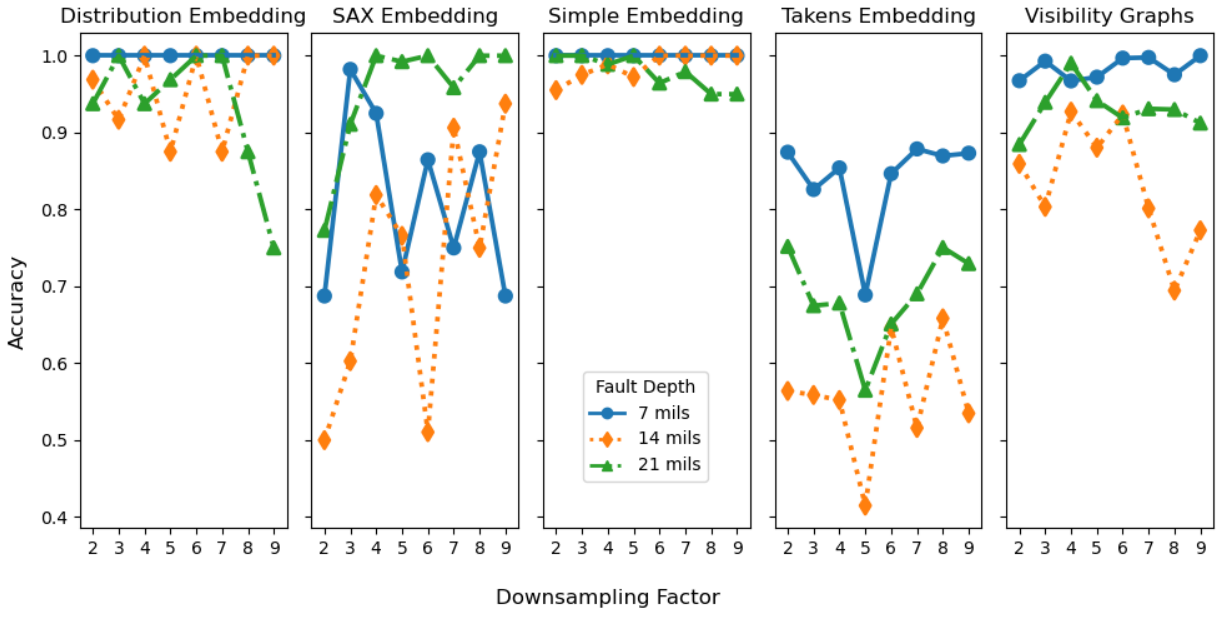


(b) Model test Performance.

Figure 3. Model performance on downsampled data at load 0 HP.



(a) Model cross-validation performance.



(b) Model test performance.

Figure 4. Model performance on downsampled data at load 3 HP.

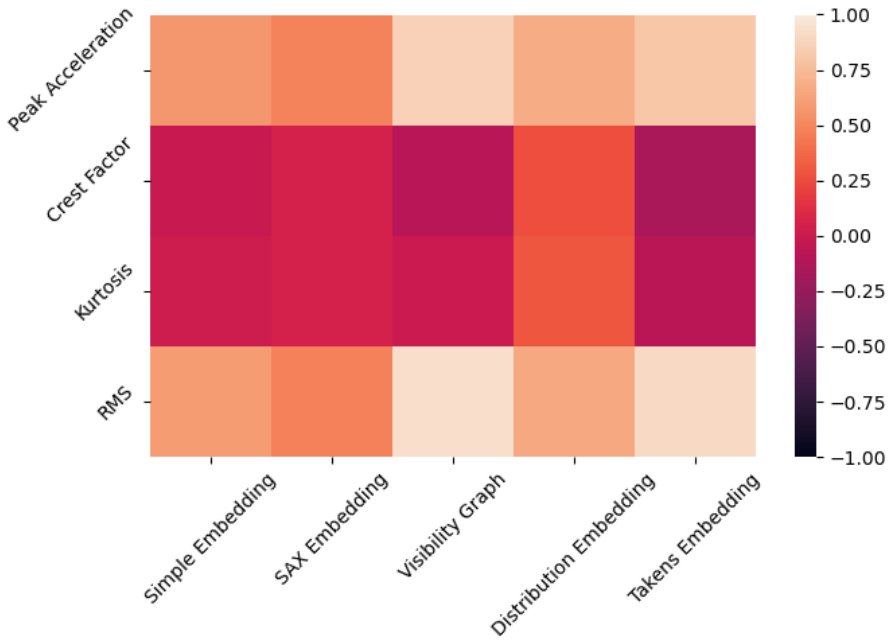


Figure 5. Correlation matrix between vibration metrics and homology group 0 persistence silhouette generated from various embeddings for the drive-end bearing fault vibration data.

Embedding	Topological Features	Homology Group	Vibration Metric
Simple Embedding	Silhouette, Landscape	0	RMS
SAX Embedding	Silhouette, Landscape, Betti	0, 1	RMS, Kurtosis
Visibility Graph	Silhouette, Landscape, Betti	0	RMS, Peak Acc., Kurtosis, Crest Factor
Distribution Embedding	Silhouette, Landscape, Betti	0, 1	RMS, Peak Acc., Kurtosis, Crest Factor
Takens Embedding	Silhouette, Landscape, Betti	0, 1	RMS, Peak Acc., Kurtosis

Table 2. Correlated topological features and vibration metrics at window size 1,000. The Homology Group column indicates whether the correlated topological features come from homology group 0 or homology group 1.

Finally, we note that the topological features from the different homology groups do not seem to have a direct connection to the vibration metrics: while homology group 1 captures higher order topological structures than homology group 0, the correlations are not necessarily restricted to the higher order vibration metrics, such as crest factor.

7. DISCUSSION

In general, the topological features provide good class separation for fault classification. The simple embedding and the distribution embedding perform well across all loads and fault conditions. On the other hand, we see that the SAX embedding generate features which lead to uneven performance for the base random forest classifier. We also examined the visibility graph as an embedding, and found that its performance is not comparable to that of the simple and distribution embeddings. These differences in model performance are exacerbated by the downsampling of the data.

We note that SAX embedding does not seem to be an ideal topological feature generator for fault classification. SAX is originally designed to detect anomalies within a single time series; the poor classifier performance seems to indicate that the vibrations encoded by SAX for a particular measurement does not generalize to other measurements.

We also investigated the potential correlations between the topological features and classical vibration metrics. While these correlations are sensitive to the choice of window sizes, nevertheless the strong correlation in some cases to known vibration metrics suggest that topological data analyses may be used to automate the vibration metric selection, which is usually done based on human intuition. Furthermore, given the good performance of the simple and distribution embeddings, and the lack of correlations between their topological features and the vibration metrics only suggests that the topological features may be better for fault diagnosis than some of the vibration metrics.

7.1 CONNECTION TO NONLINEAR DYNAMICS

The 0-1 Test [27] was developed to detect chaotic behaviors. In particular, a time series data is projected onto the p-q plane with the following equations:

$$p(n) = \sum_{j=1}^{n} \phi(j) \cos(jc),$$

and

$$q(n) = \sum_{j=1}^{n} \phi(j) \sin(jc),$$

where $\phi(j)$ is the value of the time series indexed by j, and c is a random variable with distribution $U(0, 2\pi)$.

Templeman et al. [28] used the p-q projections and created kernel density estimates, from which the following persistence metric is calculated to determine the existence of chaotic behavior:

$$PS_1 = \left\langle \sum_{j=1}^{N_i} \left\{ \frac{\sqrt{d_j^2 + b_j^2}}{N_i}, (b_j, d_j) \in D_i \right\} \right\rangle,$$

where b_j and d_j correspond to the birth and death times, respectively, of the jth structure in the ith persistence diagram D_i.

We investigate the feasibility of using p-q projections as an embedding, and determine whether the embedding could be used to capture the transition from chaotic to periodic behaviors. In this preliminary analysis, we use the Center for Intelligent Maintenance Systems (IMS), University of Cincinnati data set [29], which is made up of vibration data collected from run-to-failure bearing experiments.

It is noted in [28] that periodic behaviors correspond to PS_1 scores greater than 1. In Figure 6, the PS_1 score of a healthy bearing is shown, and note that scores never cross the PS_1 threshold. In Figure 7, the scores of a bearing that eventually develops a fault is shown; note that the PS_1 scores crosses the threshold several times prior to the recorded bearing failure.

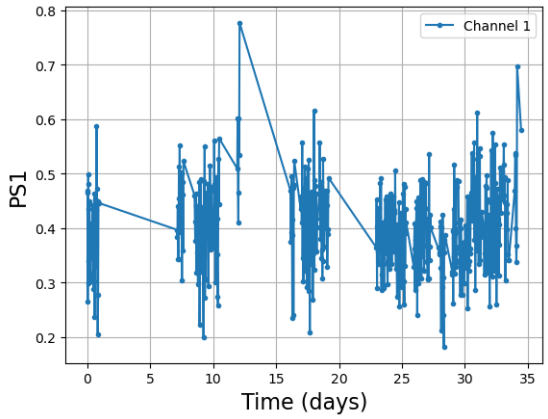


Figure 6. PS_1 scores of a bearing where a fault was never developed.

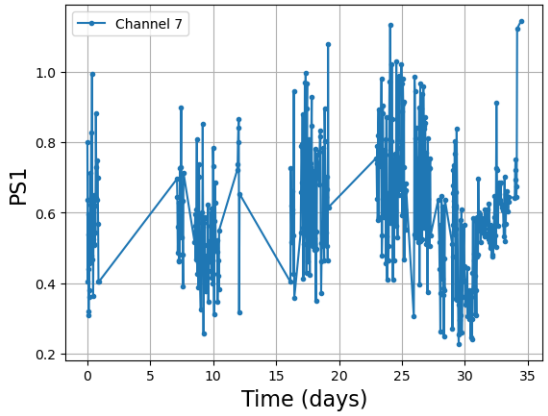


Figure 7. PS_1 scores of a bearing where a fault was eventually developed.

8. CONCLUSION

In this paper, we examined different embeddings for vibration signals, from which topological features can be generated. Furthermore, we used these topological features to build models to classify faults with the CWRU data set, across various conditions, as well as artificially downsampling the data, to examine the performance of these features for fault analysis. It was found that while some embeddings generate features which perform well, even in degraded conditions, other embeddings suffer from poor model performance.

The potential correlations of the topological features to some of the classical time-domain vibration metrics are also examined. It was found that some embeddings have features that have consistently strong correlations to the vibration metrics, suggesting that these embeddings can automatically generate these vibration metrics.

Additionally, we performed an investigation of an emerging method from the field of nonlinear dynamics that quantifies chaotic and periodic behaviors through TDA. Our preliminary analysis showed that the method could distinguish between healthy and faulty bearings; moreover, the method may provide leading indicators for failure events.

We believe that TDA is a valuable addition to the world of vibration analysis. While we have shown in this paper that some topological features provide good class separability for fault classifiers, there may be yet undiscovered topological features that are tuned specifically for vibration analyses that yield even better performance. Moreover, the topological features investigated in this paper are based on Vietoris-Rips method of simplicial complex construction; the investigation of other simplicial complexes and additional topological features for vibration analysis remains an item for future work.

This page is intentionally blank.

REFERENCES

1. Randall, R. 2011. *Vibration-based Condition Monitoring: Industrial, Aerospace and Automotive Applications*, Wiley.

2. Taylor, J. I. 2003. *The Vibration Analysis Handbook*, Vibration Consultants, 2 ed.

3. Mendel, E., Rauber, T. W., Varejão, F. M., and Batista, R. J. 2009. "Rolling element bearing fault diagnosis in rotating machines of oil extraction rigs," *2009 17th European Signal Processing Conference* (pp. 1602–1606).

4. Lacey, S. 2008. "An Overview of Bearing Vibration Analysis," *Maintenance and Asset Management Journal*, vol. 23.

5. Accorsi, R., Manzini, R., Pascarella, P., Patella, M., and Sassi, S. 2017. "Data Mining and Machine Learning for Condition-based Maintenance," *Procedia Manufacturing*, vol. 11, pp. 1153–1161.

6. Kankar, P. K., Sharma, S., and Harsha, S. P. 2011. "Rolling Element Bearing Fault Diagnosis Using Wavelet Transform," *Elsevier - Neurocomputing*, vol. 74, pp. 1638–1645.

7. Tao, S., Zhang, T., Yang, J., and Wang, X. 2015. "Bearing Fault Diagnosis Method Based on Stacked Autoencoder and Softmax Regression," *Control Conference (CCC), Hangzhou, China, July 28-30, 2015* (pp. 6331–6335), IEEE.

8. Yan, W. and Yu, L. 2015. "On Accurate and Reliable Anomaly Detection for Gas Turbine Combustors: A Deep Learning Approach," *Annual Conference of The Prognostics and Health Management Society 2015*, vol. 6.

9. Chen, Z., Li, C., and Sanchez, R.-V. 2015. "Gearbox Fault Identification and Classification with Convolutional Neural Networks," *Shock and Vibration*, vol. 2015, pp. 1–10.

10. Chazal, F. and Michel, B. 2021. "An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists," .

11. Lee, D. and Sabater, A. 2022. "Visibility Graphs, Persistent Homology, and Rolling Element Bearing Fault Detection," *2022 IEEE International Conference on Prognostics and Health Management*.

12. Lacasa, L., Luque, B., Ballesteros, F. J., Luque, J., and Nuño, J. C. 2008. "From time series to complex networks: The visibility graph," *Proceedings of the National Academy of Sciences*, vol. 105, pp. 4972 – 4975.

13. Carlsson, G. 2009. "Topology and Data," *Bulletin of The American Mathematical Society - BULL AMER MATH SOC*, vol. 46, pp. 255–308.

14. Edelsbrunner, H. and Harer, J. 2010. *Computational Topology - an Introduction.*, American Mathematical Society.

15. Edelsbrunner, H. and Harer, J. 2008. *Persistent homology - A survey*, vol. 453.

16. Chazal, F., Fasy, B. T., Lecci, F., Rinaldo, A., and Wasserman, L. 2013. "Stochastic Convergence of Persistence Landscapes and Silhouettes," .

17. Bonis, T., Ovsjanikov, M., Oudot, S., and Chazal, F. 2016. "Persistence-based pooling for shape pose recognition," *International workshop on computational topology in image context* (pp. 19–29). Springer.

18. Reininghaus, J., Huber, S., Bauer, U., and Kwitt, R. 2015. "A stable multi-scale kernel for topological machine learning," *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4741–4748).

19. Atienza, N., Gonzalez-Diaz, R., and Rucco, M. 2017. "Persistent Entropy for Separating Topological Features from Noise in Vietoris-Rips Complexes," .

20. Tauzin, G., Lupo, U., Tunstall, L., Pérez, J. B., Caorsi, M., Medina-Mardones, A. M., Dassatti, A., and Hess, K. 2021. "giotto-tda: A Topological Data Analysis Toolkit for Machine Learning and Data Exploration," *Journal of Machine Learning Research*, vol. 22, no. 39, pp. 1–6, URL `http://jmlr.org/papers/v22/20-325.html`.

21. Aggarwal, C. C., Hinneburg, A., and Keim, D. A. 2001. "On the Surprising Behavior of Distance Metrics in High Dimensional Spaces," J. V. den Bussche and V. Vianu, eds., *Database Theory - ICDT 2001, 8th International Conference, London, UK, January 4-6, 2001, Proceedings*, *Lecture Notes in Computer Science*, vol. 1973 (pp. 420–434), Springer.

22. Cha, S.-H. 2007. "Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions," *International Journal of Mathematical Models and Methods in Applied Sciences*, vol. 1, no. 4, pp. 300–307.

23. Bergillos, C. "ts2vg: Time series to visibility graphs," `https://cbergillos.com/ts2vg/`.

24. Takens, F. 1981. "Detecting Strange Attractors in Turbulence," D. Rand and L.-S. Young, eds., *Dynamical Systems and Turbulence, Warwick 1980* (pp. 366–381), Springer Berlin Heidelberg.

25. Keogh, E., Lin, J., and Fu, A. 2005. "HOT SAX: Efficiently finding the most unusual time series subsequence," *5th IEEE International Conference on Data Mining (ICDM)* (pp. 226–233).

26. Case Western Reserve University Bearing Data Center. "Seeded Fault Test Data," `https://engineering.case.edu/bearingdatacenter`.

27. Gottwald, G. A. and Melbourne, I. 2009. "On the Implementation of the 0–1 Test for Chaos," *SIAM Journal on Applied Dynamical Systems*, vol. 8, no. 1, pp. 129–145.

28. Tempelman, J. R. and Khasawneh, F. A. 2020. "A look into chaos detection through topological data analysis," *Physica D: Nonlinear Phenomena*, vol. 406, p. 132446.

29. Lee, J., Qiu, H., Yu, G., Lin, J., and Rexnord Technical Services. 2007. "Bearing Data Set," NASA Ames Prognostics Data Repository (`http://ti.arc.nasa.gov/project/prognostic-data-repository`), NASA Ames Research Center, Moffett Field, CA.

INITIAL DISTRIBUTION

84310	Technical Library/Archives	(1)
53629	D. Lee	(1)
53629	J. Rorie	(1)
71740	A. Sabater	(1)

Defense Technical Information Center
Fort Belvoir, VA 22060–6218 (1)

This page is intentionally blank.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-01-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
November 2023	Final	

4. TITLE AND SUBTITLE

An Investigation of Time Series Embeddings for Topological Data Analysis

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHORS

Dean Lee
Jamal Rorie, Ph.D.
Andrew Sabater, Ph.D.
NIWC Pacific

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

NIWC Pacific
53560 Hull Street
San Diego, CA 92152–5001

8. PERFORMING ORGANIZATION REPORT NUMBER

TR-3323

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

N/A

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Distribution A: Approved for public release. Distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Topological data analysis (TDA) is an important part of a data scientist's tool box; through the extraction of topological information, TDA provides an automated means of feature engineering. Thus, TDA may be an appropriate tool for fault analysis of mechanical systems, where the methodologies have traditionally relied on expert intuition and signals processing techniques. However, appropriate embedding of the data must be defined before TDA can be applied. In this report, we investigate several different embeddings of time series data for the application of TDA for fault analysis, and show that features engineered by TDA improve fault classification accuracy.

15. SUBJECT TERMS

Topological data analysis; machine learning; data analytics; data science

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Dean Lee
U	U	U	SAR	34	**19B. TELEPHONE NUMBER** *(Include area code)* (619) 553-7203

This page is intentionally blank.

This page is intentionally blank.

Naval Information Warfare Center (NIWC) Pacific
San Diego, CA 92152-5001