Time series classification at scale

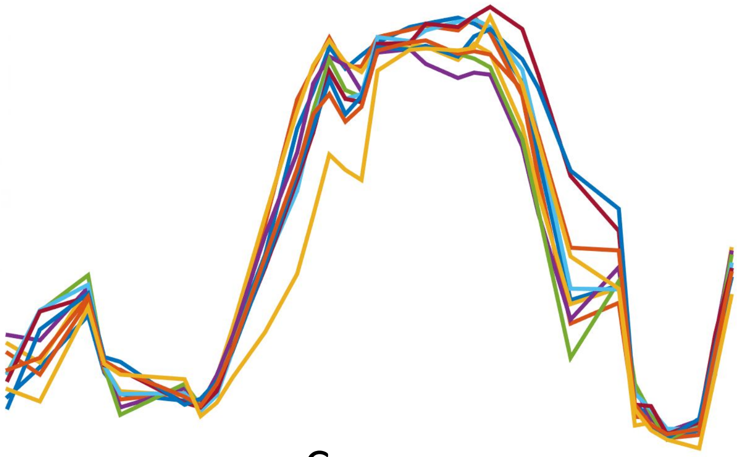
François Petitjean

http://francois-petitjean.com

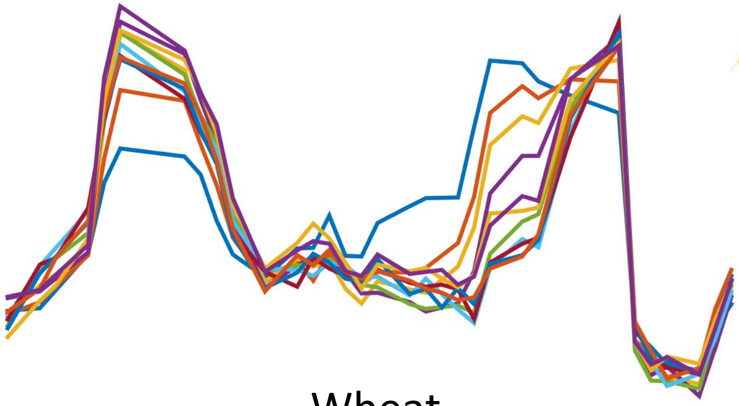
PhD work of A. Shifaz, B. Lucas and H. Ismail Fawaz

… with G. Forestier, C. Pelletier, G. Webb, J. Weber, B. Goethals, J. Weber, D. Schmidt, L. Idoumghar, P. Muller, L. O’Neill, and N. Zaidi

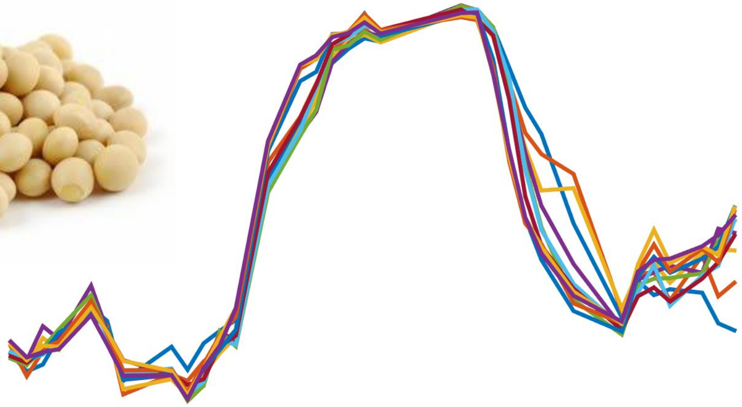
Time series classification



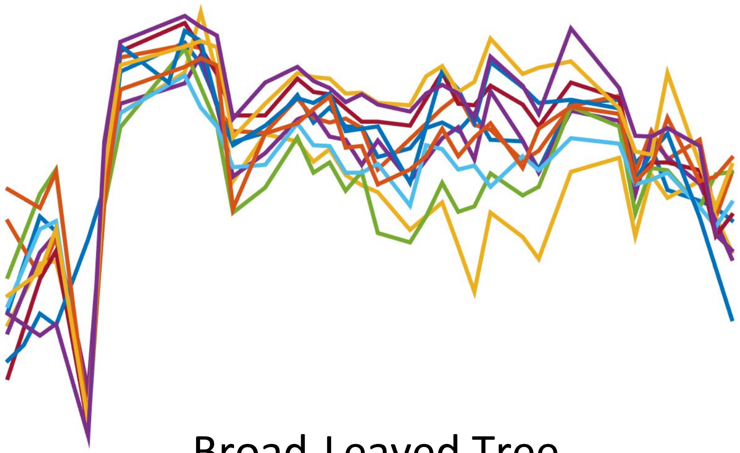
Corn



Wheat



Soybean

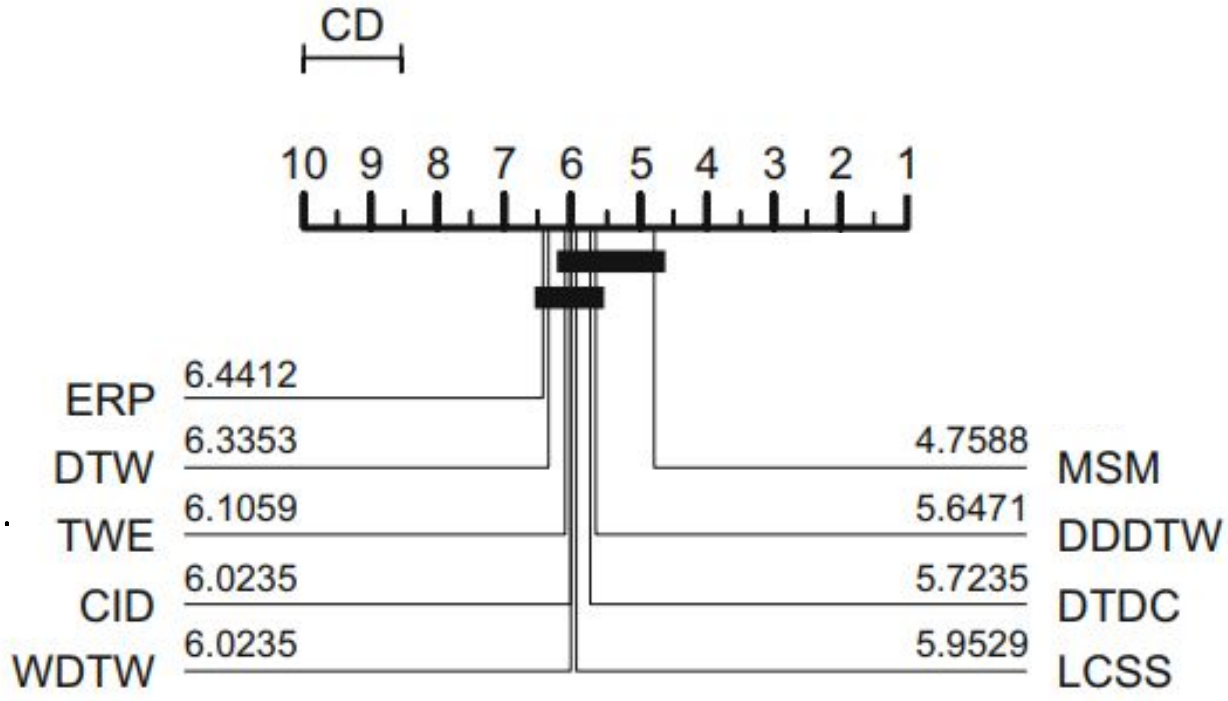


Broad-Leaved Tree

Research into time series classification has accelerated very quickly over the last 5 years

Until recently,

- Many specialised time series classifiers developed
- But none dominated on accuracy on the UCR repository (85 datasets)

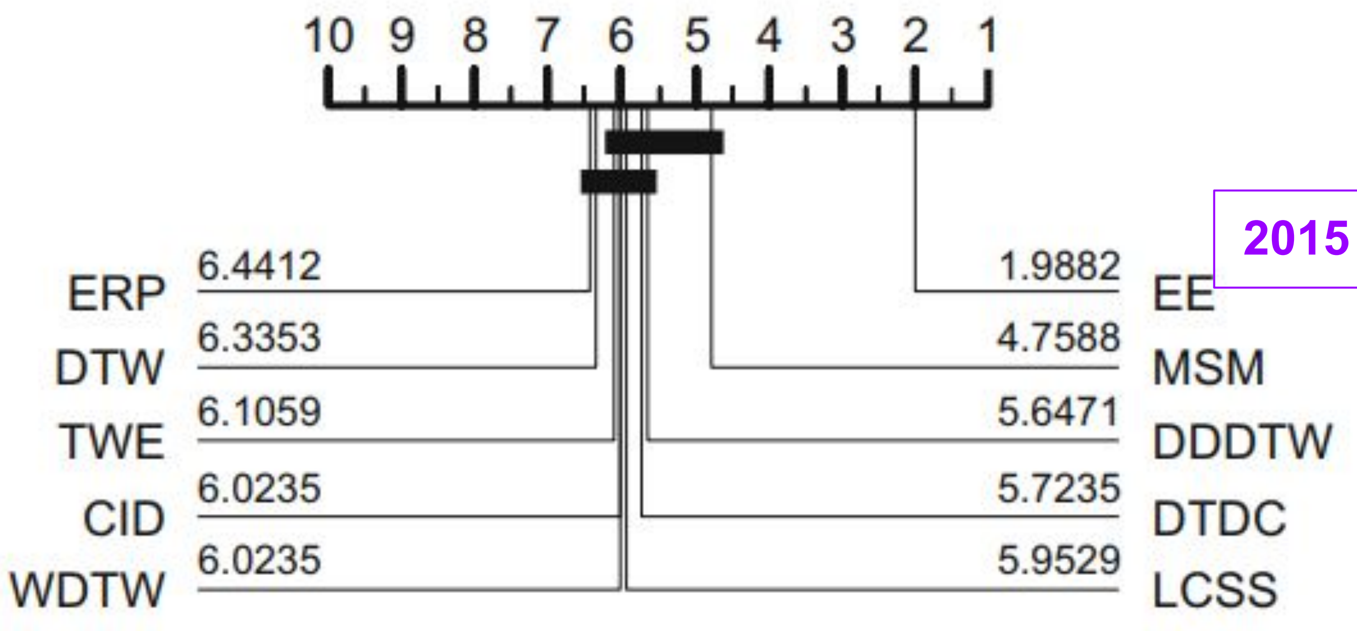


Bagnall, A., Lines, J., Bostrom, A., Large, J., & Keogh, E. (2017). The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, *31*(3), 606-660.

A revolution in time series classification

Ensembles* have swept all before them!

* i.e. Tony, Jason, James and Aaron :)



2015

Data Min Knowl Disc
DOI 10.1007/s10618-014-0361-2

Time series classification with ensembles of elastic distance measures

Jason Lines · Anthony Bagnall

Received: 10 January 2014 / Accepted: 5 June 2014
© The Author(s) 2014

Abstract Several alternative distance measures for comparing time series have recently been proposed and evaluated on time series classification (TSC) problems. These include variants of dynamic time warping (DTW), such as weighted and derivative DTW, and edit distance-based measures, including longest common subsequence, edit distance with real penalty, time warp with edit, and move–split–merge. These measures have the common characteristic that they operate in the time domain and compensate for potential localised misalignment through some elastic adjustment. Our aim is to experimentally test two hypotheses related to these distance measures. Firstly, we test whether there is any significant difference in accuracy for TSC problems between nearest neighbour classifiers using these distance measures. Secondly, we test whether combining these elastic distance measures through simple ensemble schemes gives significantly better accuracy. We test these hypotheses by carrying out one of the largest experimental studies ever conducted into time series classification. Our first key finding is that there is no significant difference between the elastic distance measures in terms of classification accuracy on our data sets. Our second finding, and the major contribution of this work, is to define an ensemble classifier that significantly outperforms the individual classifiers. We also demonstrate that the ensemble is more accurate than approaches not based in the time domain. Nearly all TSC papers in the data mining literature cite DTW (with warping window set through cross validation) as the benchmark for comparison. We believe that our ensemble is the first ever classifier to significantly outperform DTW and as such raises the bar for future work in this area.

Responsible editor: Eamonn Keogh.

J. Lines (✉) · A. Bagnall
University of East Anglia, Norwich, UK
e-mail: j.lines@uea.ac.uk

A. Bagnall
e-mail: anthony.bagnall@uea.ac.uk

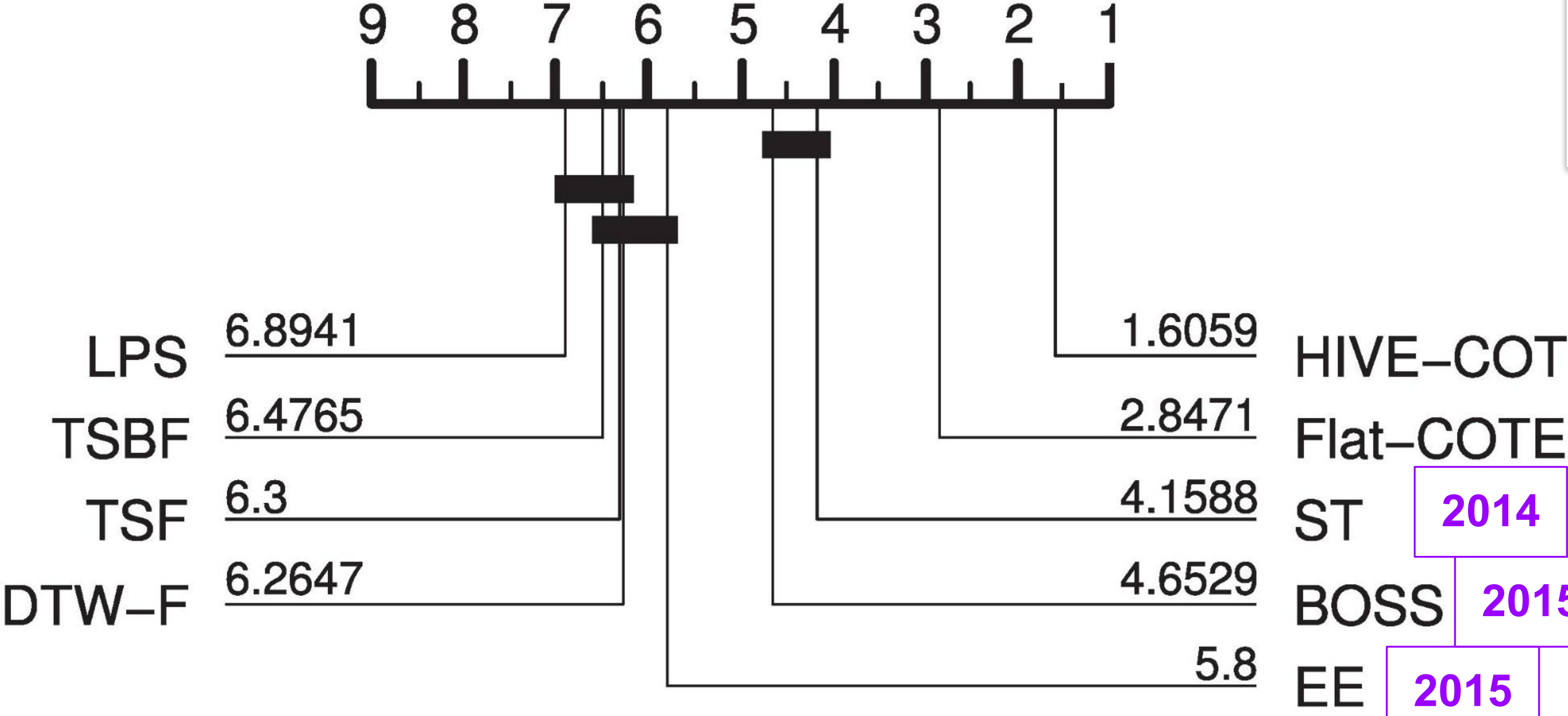
Published online: 28 June 2014

Springer

Lines, J. & Bagnall, A., Time Series Classification with Ensembles of Elastic Distance Measures, *Data Mining and Knowledge Discovery*, 2015.

A revolution in time series classification

They didn’t stop there: a leap forward around 2015



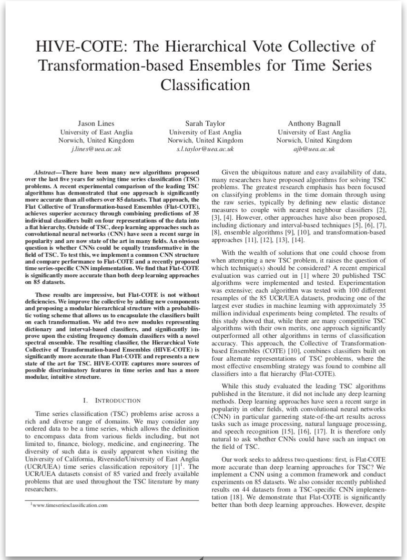
2016

2015

2014

2015

2015



HIVE-COTE: The Hierarchical Vote Collective of Transformation-based Ensembles for Time Series Classification

Jason Lines
University of East Anglia
Norwich, United Kingdom
j.lines@uea.ac.uk

Sarah Taylor
University of East Anglia
Norwich, United Kingdom
s.l.taylor@uea.ac.uk

Anthony Bagnall
University of East Anglia
Norwich, United Kingdom
ajb@uea.ac.uk

***Abstract*—There have been many new algorithms proposed over the last five years for solving time series classification (TSC) problems. A recent experimental comparison of the leading TSC algorithms has demonstrated that one approach is significantly more accurate than all others over 85 datasets. That approach, the Flat Collective of Transformation-based Ensembles (Flat-COTE), achieves superior accuracy through combining predictions of 35 individual classifiers built on four representations of the data into a flat hierarchy. Outside of TSC, deep learning approaches such as convolutional neural networks (CNN) have seen a recent surge in popularity and are now state of the art in many fields. An obvious question is whether CNNs could be equally transformative in the field of TSC. To test this, we implement a common CNN structure and compare performance to Flat-COTE and a recently proposed time series-specific CNN implementation. We find that Flat-COTE is significantly more accurate than both deep learning approaches on 85 datasets.**

These results are impressive, but Flat-COTE is not without deficiencies. We improve the collective by adding new components and proposing a modular hierarchical structure with a probabilistic voting scheme that allows us to encapsulate the classifiers built on each transformation. We add two new modules representing dictionary and interval-based classifiers, and significantly improve upon the existing frequency domain classifiers with a novel spectral ensemble. The resulting classifier, the Hierarchical Vote Collective of Transformation-based Ensembles (HIVE-COTE) is significantly more accurate than Flat-COTE and represents a new state of the art for TSC. HIVE-COTE captures more sources of possible discriminatory features in time series and has a more modular, intuitive structure.

I. INTRODUCTION

Time series classification (TSC) problems arise across a rich and diverse range of domains. We may consider any ordered data to be a time series, which allows the definition to encompass data from various fields including, but not limited to, finance, biology, medicine, and engineering. The diversity of such data is easily apparent when visiting the University of California, Riverside/University of East Anglia (UCR/UEA) time series classification repository [1][1]. The UCR/UEA datasets consist of 85 varied and freely available problems that are used throughout the TSC literature by many researchers.

Given the ubiquitous nature and easy availability of data, many researchers have proposed algorithms for solving TSC problems. The greatest research emphasis has been focused on classifying problems in the time domain through using the raw series, typically by defining new elastic distance measures to couple with nearest neighbour classifiers [2], [3], [4]. However, other approaches have also been proposed, including dictionary and interval-based techniques [5], [6], [7], [8], ensemble algorithms [9], [10], and transformation-based approaches [11], [12], [13], [14].

With the wealth of solutions that one could choose from when attempting a new TSC problem, it raises the question of which technique(s) should be considered? A recent empirical evaluation was carried out in [1] where 20 published TSC algorithms were implemented and tested. Experimentation was extensive; each algorithm was tested with 100 different resamples of the 85 UCR/UEA datasets, producing one of the largest ever studies in machine learning with approximately 35 million individual experiments being completed. The results of this study showed that, while there are many competitive TSC algorithms with their own merits, one approach significantly outperformed all other algorithms in terms of classification accuracy. This approach, the Collective of Transformation-based Ensembles (COTE) [10], combines classifiers built on four alternate representations of TSC problems, where the most effective ensembling strategy was found to combine all classifiers into a flat hierarchy (Flat-COTE).

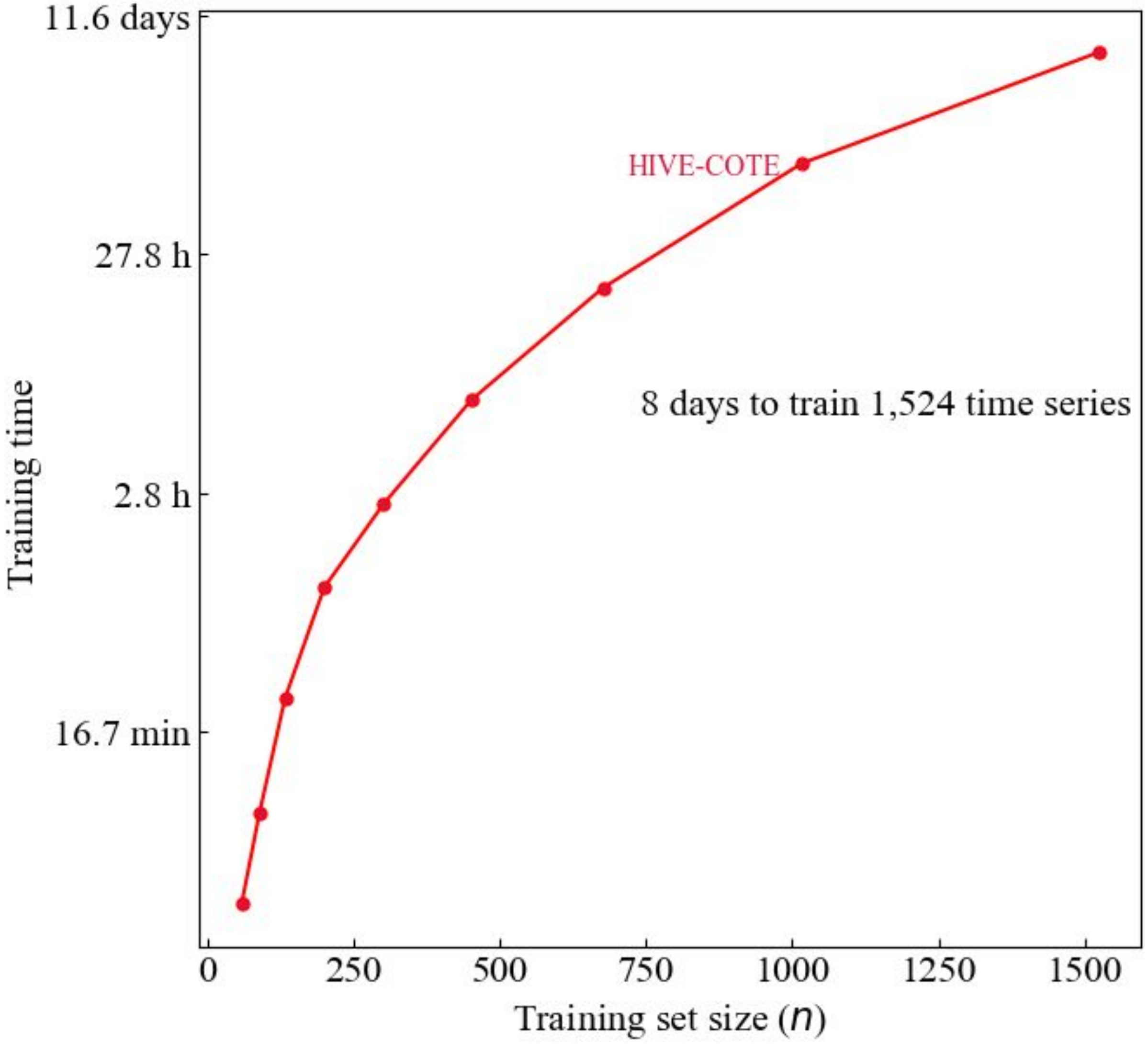
While this study evaluated the leading TSC algorithms published in the literature, it did not include any deep learning methods. Deep learning approaches have seen a recent surge in popularity in other fields, with convolutional neural networks (CNN) in particular garnering state-of-the-art results across tasks such as image processing, natural language processing, and speech recognition [15], [16], [17]. It is therefore only natural to ask whether CNNs could have such an impact on the field of TSC.

Our work seeks to address two questions: first, is Flat-COTE more accurate than deep learning approaches for TSC? We implement a CNN using a common framework and conduct experiments on 85 datasets. We also consider recently published results on 44 datasets from a TSC-specific CNN implementation [18]. We demonstrate that Flat-COTE is significantly better than both deep learning approaches. However, despite

[1] www.timeseriesclassification.com

Lines, J., Taylor, S., & Bagnall, A. (2016). Hive-cote: The hierarchical vote collective of transformation-based ensembles for time series classification. In *2016 IEEE 16th international conference on data mining (ICDM)* (pp. 1041-1046). IEEE.

However, the most accurate ensembles do not scale



State-of-the-art methods	**Training time complexity** (n training time series of length l)
Historical baseline	
1-NN with Dynamic Time Warping (DTW) [Ratanamahatana and Keogh, 2005 ; Ding *et al.*, 2008]	$O(n^2 l^3)$
window size set by cross-validation	
Four leading classification algorithms [Bagnall *et al.*, 2017]	
Bag-Of-SFA-Symbols (BOSS) [Schäfer, 2015]	$O(n^2 l^3)$
Shapelet Transform (ST) [Hills *et al.*, 2014]	$O(n^2 l^4)$
Elastic Ensembles (EE) [Lines and Bagnall, 2015]	$O(n^2 l^3)$
Collective Of Transformation-based Ensembles (COTE) [Bagnall *et al.*, 2015]	lower bounded by EE and ST algorithms

For 1M training instances, training the EE algorithm would require 73,000 days, **200 years**!

However, the most accurate ensembles do not scale

2017

Data Min Knowl Disc
DOI 10.1007/s10618-016-0483-9

CrossMark

The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances

**Anthony Bagnall[1] · Jason Lines[1] ·
Aaron Bostrom[1] · James Large[1] · Eamonn Keogh[2]**

Received: 7 June 2016 / Accepted: 1 November 2016

Abstract In the last 5 years there have been a large number of new time series classification algorithms proposed in the literature. These algorithms have been evaluated on subsets of the 47 data sets in the University of California, Riverside time series classification archive. The archive has recently been expanded to 85 data sets, over half of which have been donated by researchers at the University of East Anglia. Aspects of previous evaluations have made comparisons between algorithms difficult. For example, several different programming languages have been used, experiments involved a single train/test split and some used normalised data whilst others did not. The relaunch of the archive provides a timely opportunity to thoroughly evaluate algorithms on a larger number of datasets. We have implemented 18 recently proposed algorithms in a common Java framework and compared them against two standard benchmark classifiers (and each other) by performing 100 resampling experiments on each of the 85 datasets. We use these results to test several hypotheses relating to whether the algorithms are significantly more accurate than the benchmarks and each other. Our results indicate that only nine of these algorithms are significantly more accurate than both benchmarks and that one classifier, the collective of transformation ensembles, is significantly more accurate than all of the others. All of our experiments and results are reproducible: we release all of our code, results and experimental details and we hope these experiments form the basis for more robust testing of new algorithms in the future.

Responsible editor: Johannes Fuernkranz.

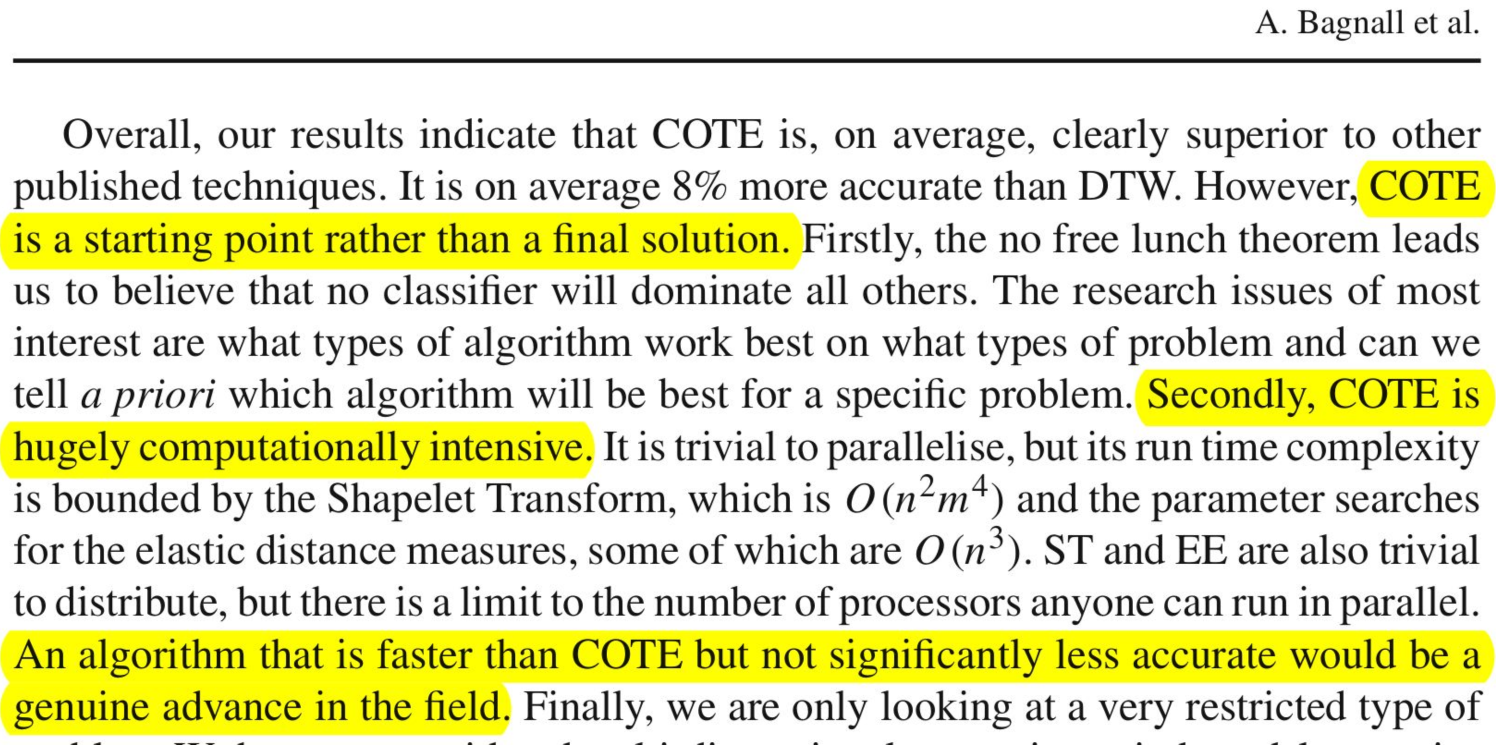
✉ Anthony Bagnall
ajb@uea.ac.uk

[1] School of Computing Sciences, University of East Anglia, Norwich, UK

[2] Computer Science & Engineering Department, University of California, Riverside, Riverside, CA, USA

Published online: 23 November 2016

Springer



Overall, our results indicate that COTE is, on average, clearly superior to other published techniques. It is on average 8% more accurate than DTW. However, COTE is a starting point rather than a final solution. Firstly, the no free lunch theorem leads us to believe that no classifier will dominate all others. The research issues of most interest are what types of algorithm work best on what types of problem and can we tell *a priori* which algorithm will be best for a specific problem. Secondly, COTE is hugely computationally intensive. It is trivial to parallelise, but its run time complexity is bounded by the Shapelet Transform, which is $O(n^2m^4)$ and the parameter searches for the elastic distance measures, some of which are $O(n^3)$. ST and EE are also trivial to distribute, but there is a limit to the number of processors anyone can run in parallel. An algorithm that is faster than COTE but not significantly less accurate would be a genuine advance in the field. Finally, we are only looking at a very restricted type of

Bagnall, A., Lines, J., Bostrom, A., Large, J., & Keogh, E. (2017). The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, *31*(3), 606-660.

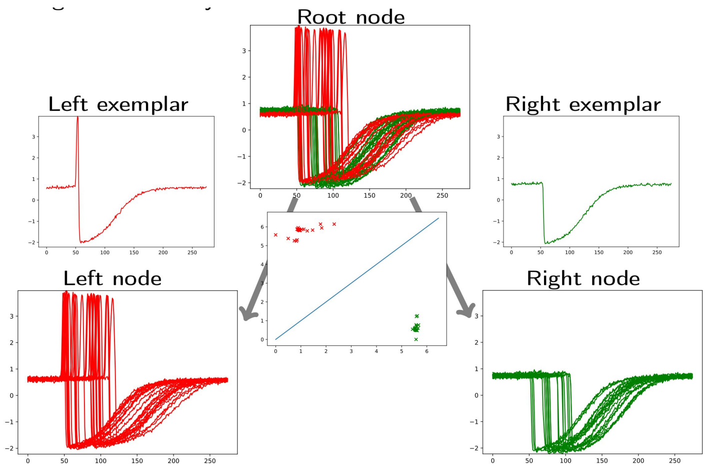
Talk Outline

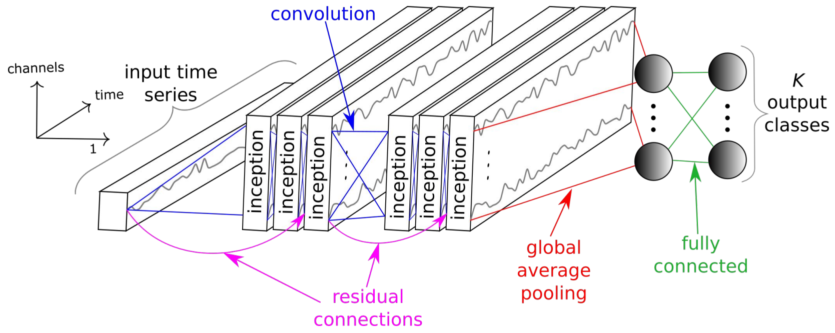
Highly accurate and scalable TSCs

- Tree-based: *Proximity Forest* and *TS-CHIEF*
- Deep Learning: *InceptionTime*

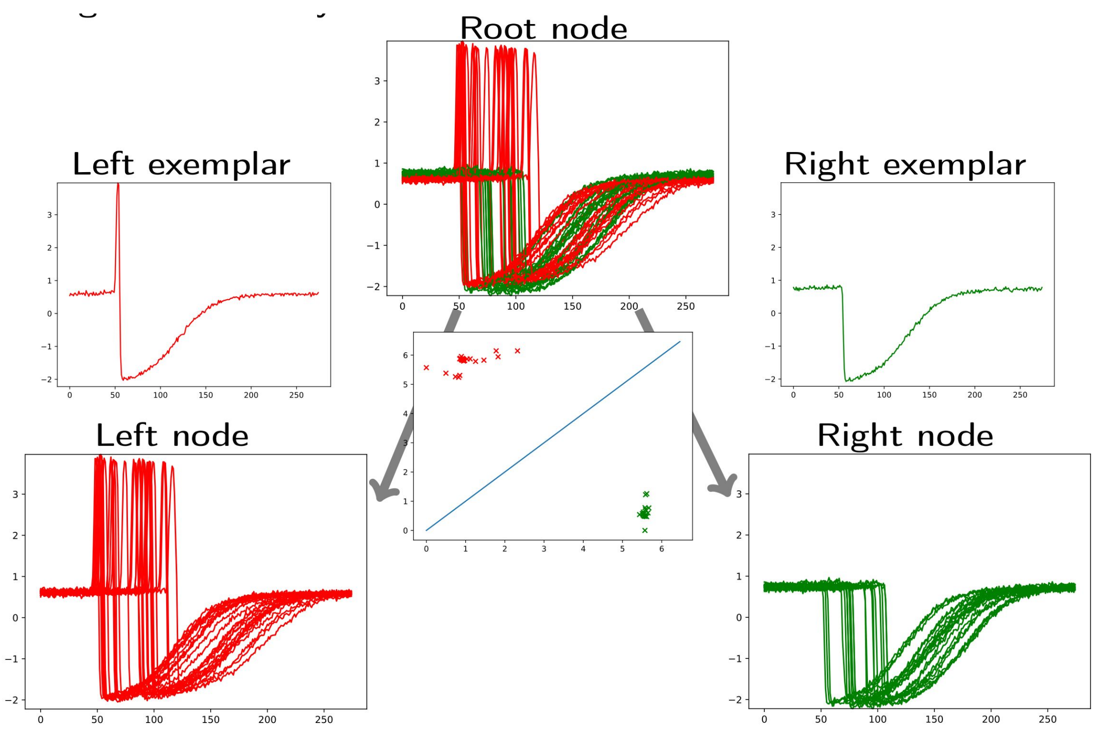
This talk is super fresh!

- 1 DAMI 2019 paper
- 2 arxiv papers submitted in the last 3 months





Part 1: Proximity Forest (PF)



2019

Data Min Knowl Disc
DOI 10.1007/s10618-016-0483-9

CrossMark

The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances

Anthony Bagnall[1] · Jason Lines[1] · Aaron Bostrom[1] · James Large[1] · Eamonn Keogh[2]

Received: 7 June 2016 / Accepted: 1 November 2016

Abstract In the last 5 years there have been a large number of new time series classification algorithms proposed in the literature. These algorithms have been evaluated on subsets of the 47 data sets in the University of California, Riverside time series classification archive. The archive has recently been expanded to 85 data sets, over half of which have been donated by researchers at the University of East Anglia. Aspects of previous evaluations have made comparisons between algorithms difficult. For example, several different programming languages have been used, experiments involved a single train/test split and some used normalised data whilst others did not. The relaunch of the archive provides a timely opportunity to thoroughly evaluate algorithms on a larger number of datasets. We have implemented 18 recently proposed algorithms in a common Java framework and compared them against two standard benchmark classifiers (and each other) by performing 100 resampling experiments on each of the 85 datasets. We use these results to test several hypotheses relating to whether the algorithms are significantly more accurate than the benchmarks and each other. Our results indicate that only nine of these algorithms are significantly more accurate than both benchmarks and that one classifier, the collective of transformation ensembles, is significantly more accurate than all of the others. All of our experiments and results are reproducible: we release all of our code, results and experimental details and we hope these experiments form the basis for more robust testing of new algorithms in the future.

Responsible editor: Johannes Fuernkranz.

✉ Anthony Bagnall
ajb@uea.ac.uk

[1] School of Computing Sciences, University of East Anglia, Norwich, UK

[2] Computer Science & Engineering Department, University of California, Riverside, Riverside, CA, USA

Published online: 23 November 2016

Springer

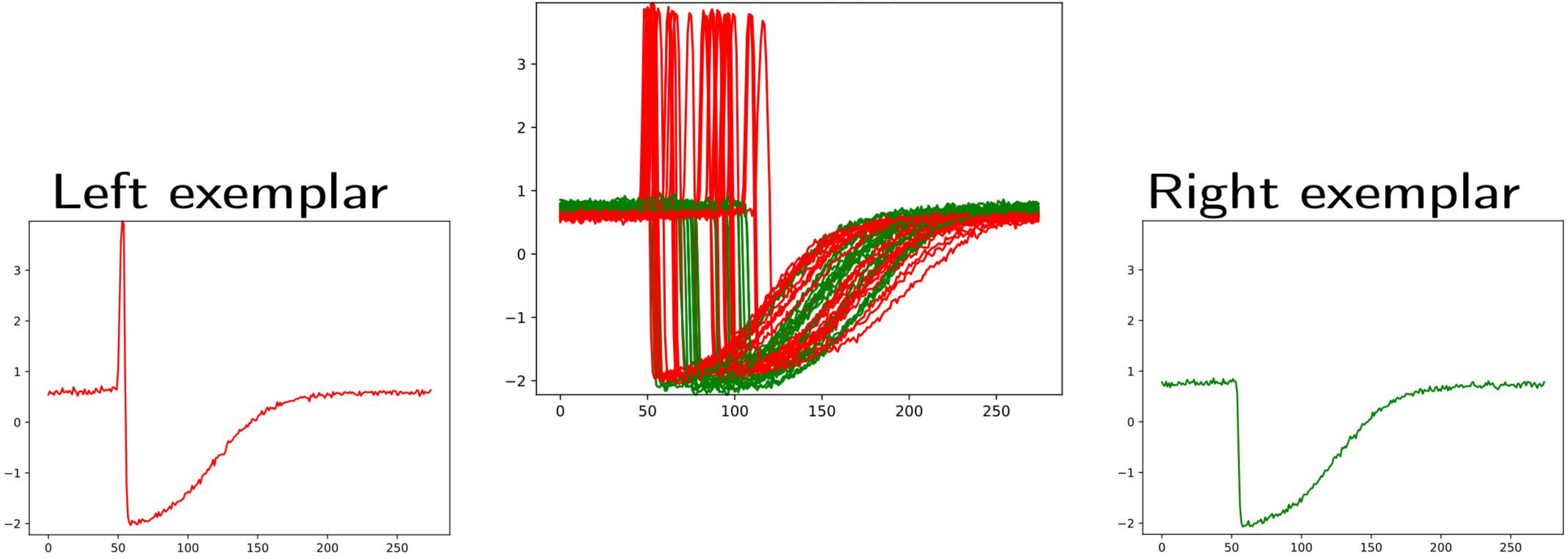
http://bit.ly/ProximityForest

B. Lucas, A. Shifaz, C. Pelletier, L. O'Neill, N. Zaidi, B. Goethals, F. Petitjean, G. Webb (2019). Proximity Forest: An effective and scalable distance-based classifier for time series. *Data Mining and Knowledge Discovery*, *33*(3), 607-635.

Proximity Forest

Starting point: How to make Elastic Ensemble (EE) scalable?

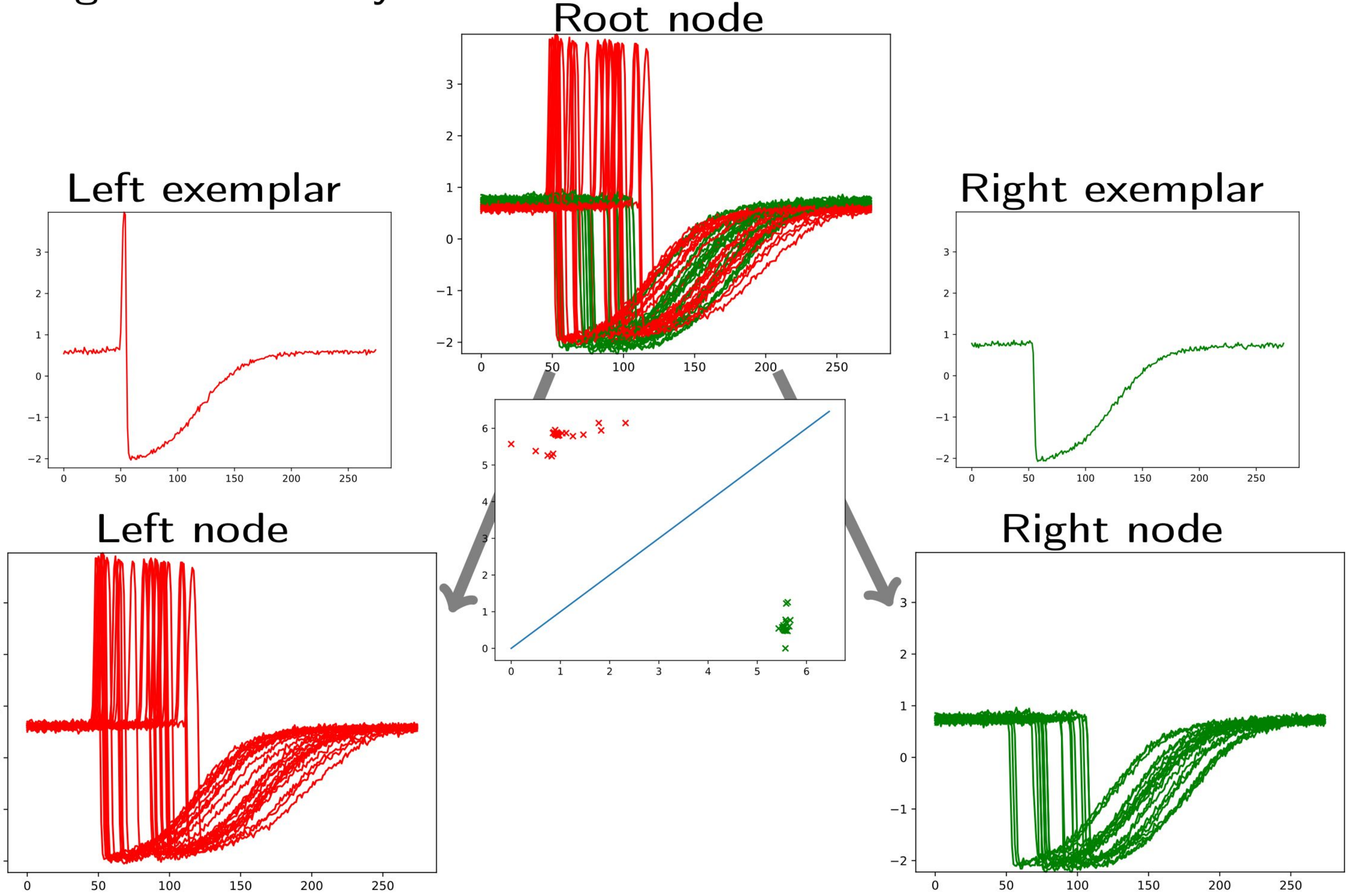
- We need a divide-and-conquer approach to be efficient
- We want to emulate Elastic Ensemble as closely as possible to allow clear comparison of fundamental strategies
- But tree-base splits don't work for time series because no attribute/value representation



Left exemplar

Root node

Right exemplar



Root node

Left exemplar

Right exemplar

Left node

Right node

Proximity Tree

- Replace conventional decision tree splits with similarity comparisons using specialised time series methods

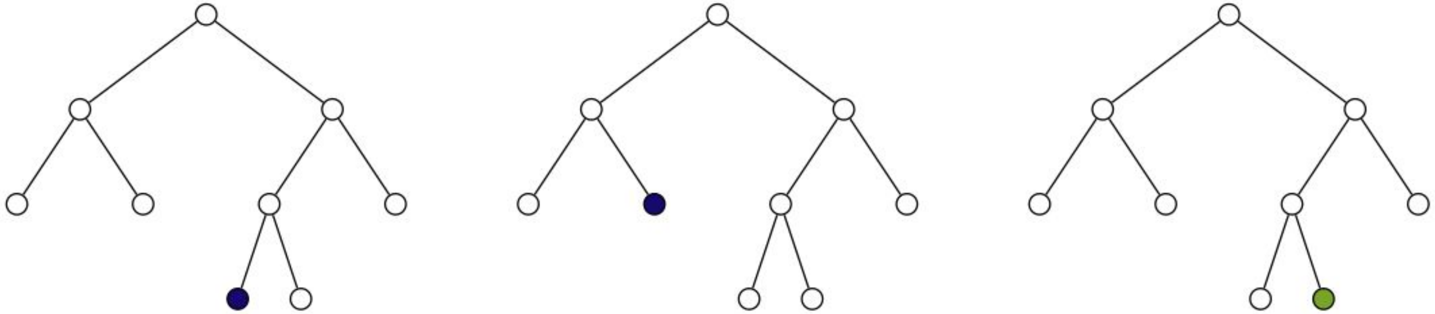
  → Makes the most of 40 years of research into designing appropriate measures for time series (DTW, TWE, MSM, LCSS, etc)

- Each branch has an exemplar associated with it

- One exemplar per class

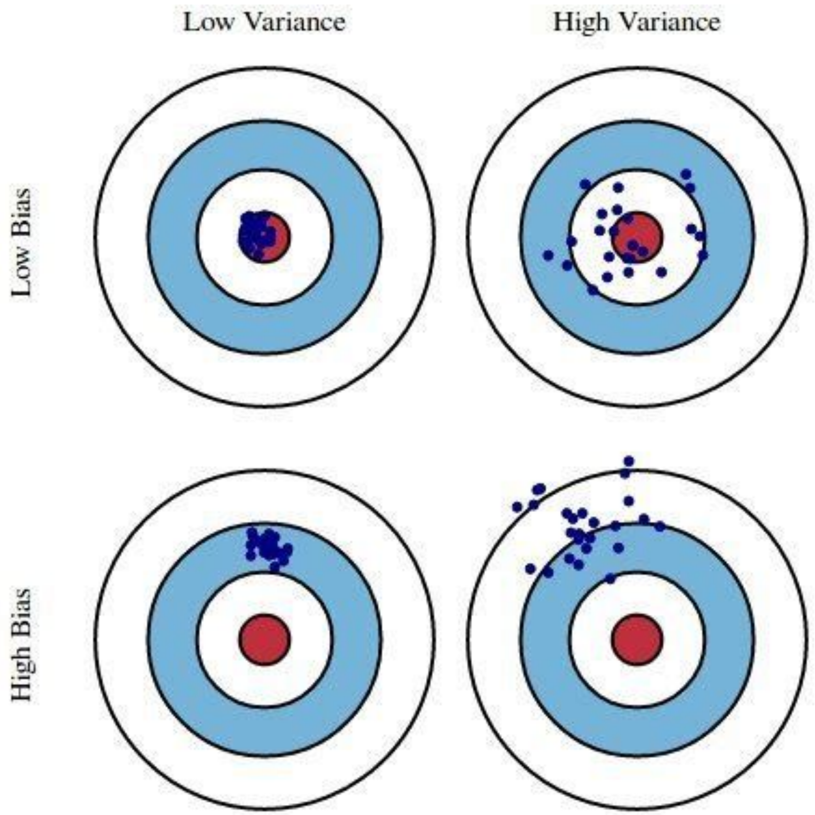
- Each split in the tree has (1) a measure and (2) a parametrization

- For classification, series *S* p most similar



Stochastic choices for speed and diversity

- Exemplars chosen at random among series at the node
- Distance measures and their parameterizations chosen at random from those used by EE
- Random choices have low bias and ensembling removes the resulting variance
- The major training time cost is passing training examples down the tree



$$\mathrm{MSE}(H) = \overline{bias}(H)^2 + \frac{1}{|H|}\overline{variance}(H) + \left(1 - \frac{1}{|H|}\right)\overline{covariance}(H)$$

N. Ueda and R. Nakano, "Generalization error of ensemble estimators," *Proceedings of International Conference on Neural Networks (ICNN'96)*, Washington, DC, USA, 1996, pp. 90-95 vol.1. doi: 10.1109/ICNN.1996.548872

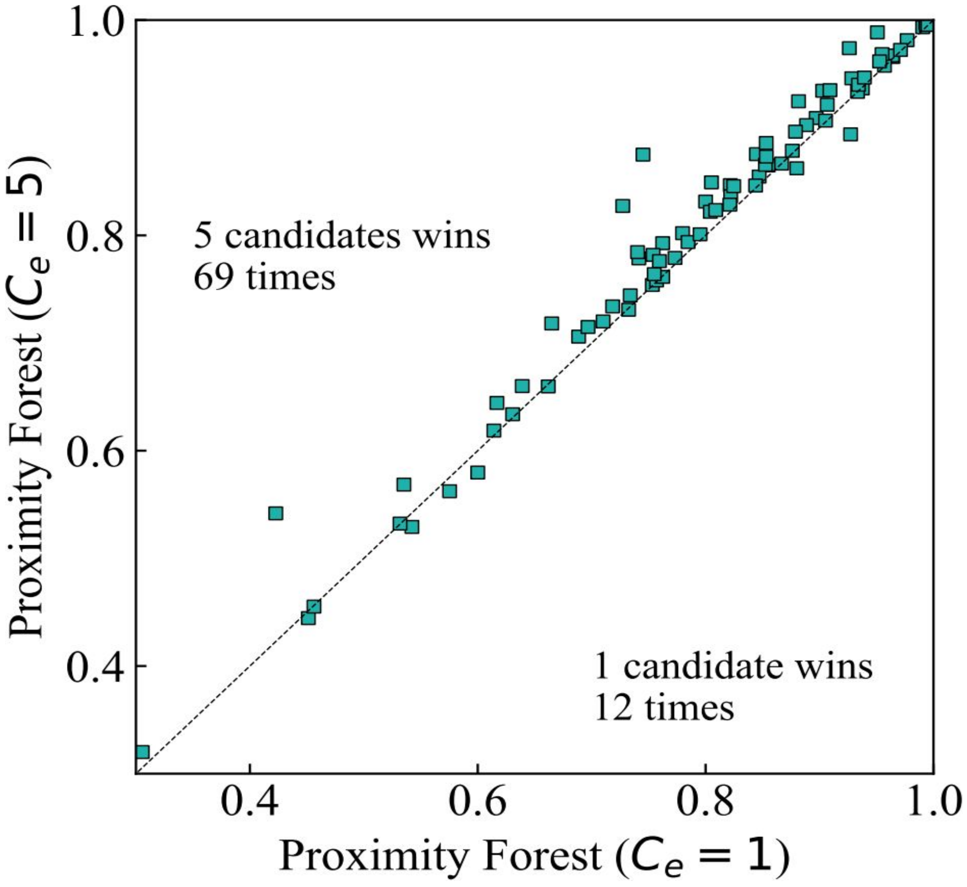
Select between multiple random candidates at each node

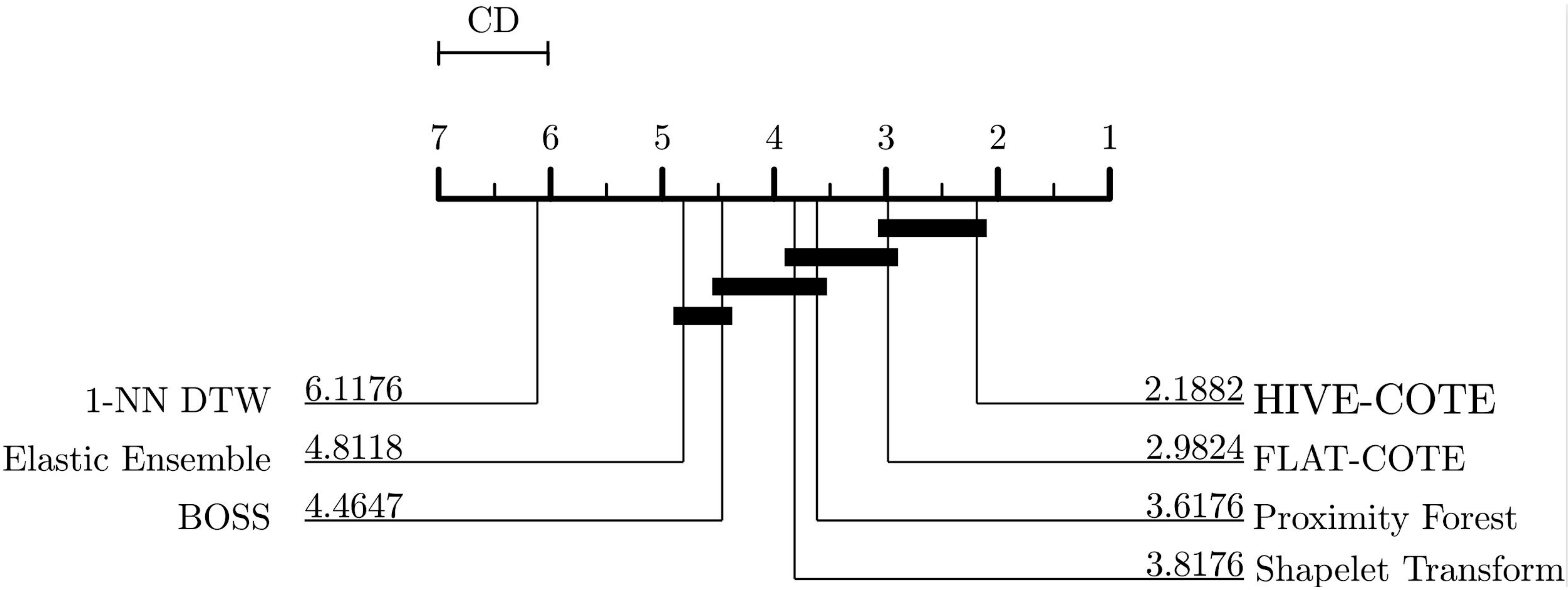
- Use GINI to select best from five candidate splits
- Increases covariance, decreases variance

→ so we don't need too many trees

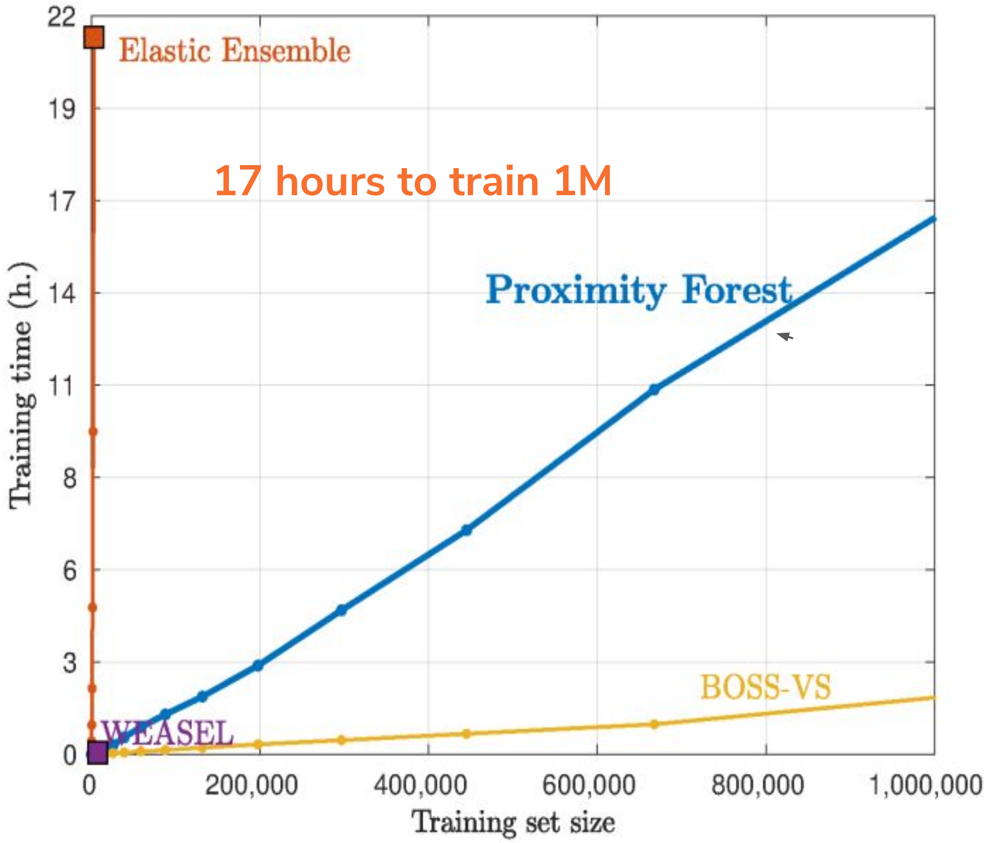
→ faster training

→ faster classification

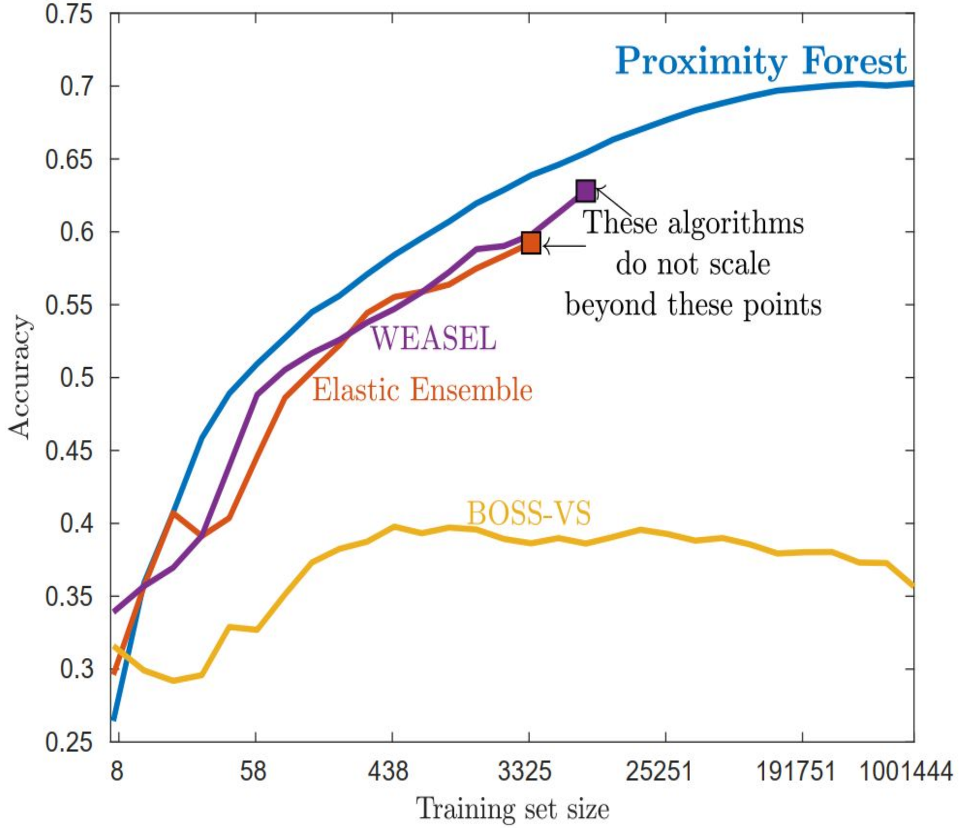



CD
7
6
5
4
3
2
1
1-NN DTW 6.1176
Elastic Ensemble 4.8118
BOSS 4.4647
2.1882 HIVE-COTE
2.9824 FLAT-COTE
3.6176 Proximity Forest
3.8176 Shapelet Transform

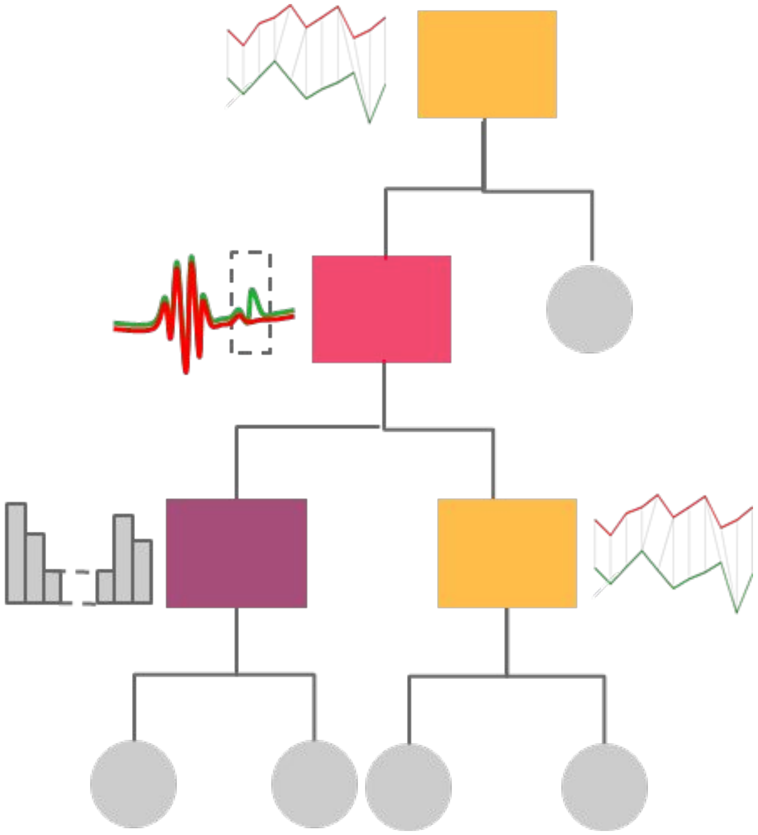
Scalability evaluated on 1M instances of Satellite Image Time Series (SITS) dataset



17 hours to train 1M



Part 2: TS-CHIEF



2019 - arxiv

Noname manuscript No.
(will be inserted by the editor)

TS-CHIEF: A Scalable and Accurate Forest Algorithm for Time Series Classification

Ahmed Shifaz · Charlotte Pelletier · François Petitjean · Geoffrey I. Webb

the date of receipt and acceptance should be inserted later

arXiv:1906.10329v1 [cs.LG] 25 Jun 2019

Abstract Time Series Classification (TSC) has seen enormous progress over the last two decades. HIVE-COTE (Hierarchical Vote Collective of Transformation-based Ensembles) is the current state of the art in terms of classification accuracy. HIVE-COTE recognizes that time series are a specific data type for which the traditional attribute-value representation, used predominantly in machine learning, fails to provide a relevant representation. HIVE-COTE combines multiple types of classifiers: each extracting information about a specific aspect of a time series, be it in the time domain, frequency domain or summarization of intervals within the series. However, HIVE-COTE (and its predecessor, FLAT-COTE) is often infeasible to run on even modest amounts of data. For instance, training HIVE-COTE on a dataset with only 1,500 time series can require 8 days of CPU time. It has polynomial runtime w.r.t training set size, so this problem compounds as data quantity increases. We propose a novel TSC algorithm, TS-CHIEF, which is highly competitive to HIVE-COTE in accuracy, but requires only a fraction of the runtime. TS-CHIEF constructs an ensemble classifier that integrates the most effective embeddings of time series that research has developed in the last decade. It uses tree-structured classifiers to do so efficiently. We assess TS-CHIEF on 85 datasets of the UCR archive, where it achieves state-of-the-art accuracy with scalability and efficiency. We demonstrate that TS-CHIEF can be trained on 130k time series in 2 days, a data quantity that is beyond the reach of any TSC algorithm with comparable accuracy.

Keywords time series, classification, metrics, bag of words, transformation, forest, scalable

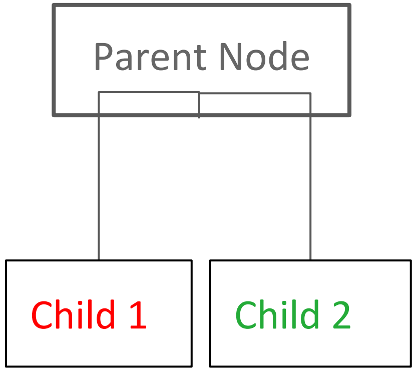
Faculty of Information Technology
25 Exhibition Walk
Monash University, Melbourne
VIC 3800, Australia
E-mail: {ahmed.shifaz,charlotte.pelletier}@monash.edu
E-mail: {francois.petitjean,geoff.webb}@monash.edu

http://bit.ly/TS-CHIEF

A. Shifaz, C. Pelletier, F. Petitjean and G. Webb (2019). TS-CHIEF: A Scalable and Accurate Forest Algorithm for Time Series Classification. *under review.* *https://arxiv.org/abs/1906.10329*

Time Series Combination of Integrated Embeddings Forest (TS-CHIEF)

Similarity-based (Proximity Forest)



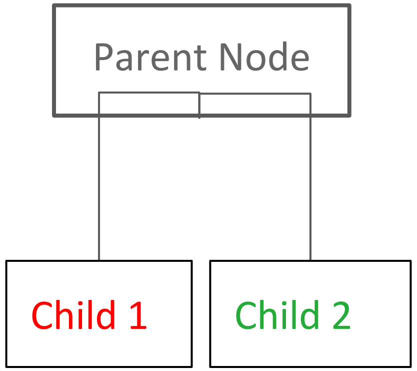
Parent Node

Child 1

Child 2

Candidate split 1

Dictionary-based



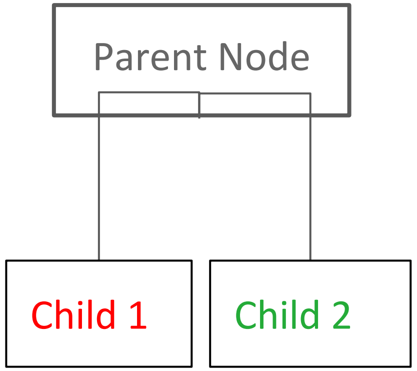
Parent Node

Child 1

Child 2

Candidate split 2

Interval-based



Parent Node

Child 1

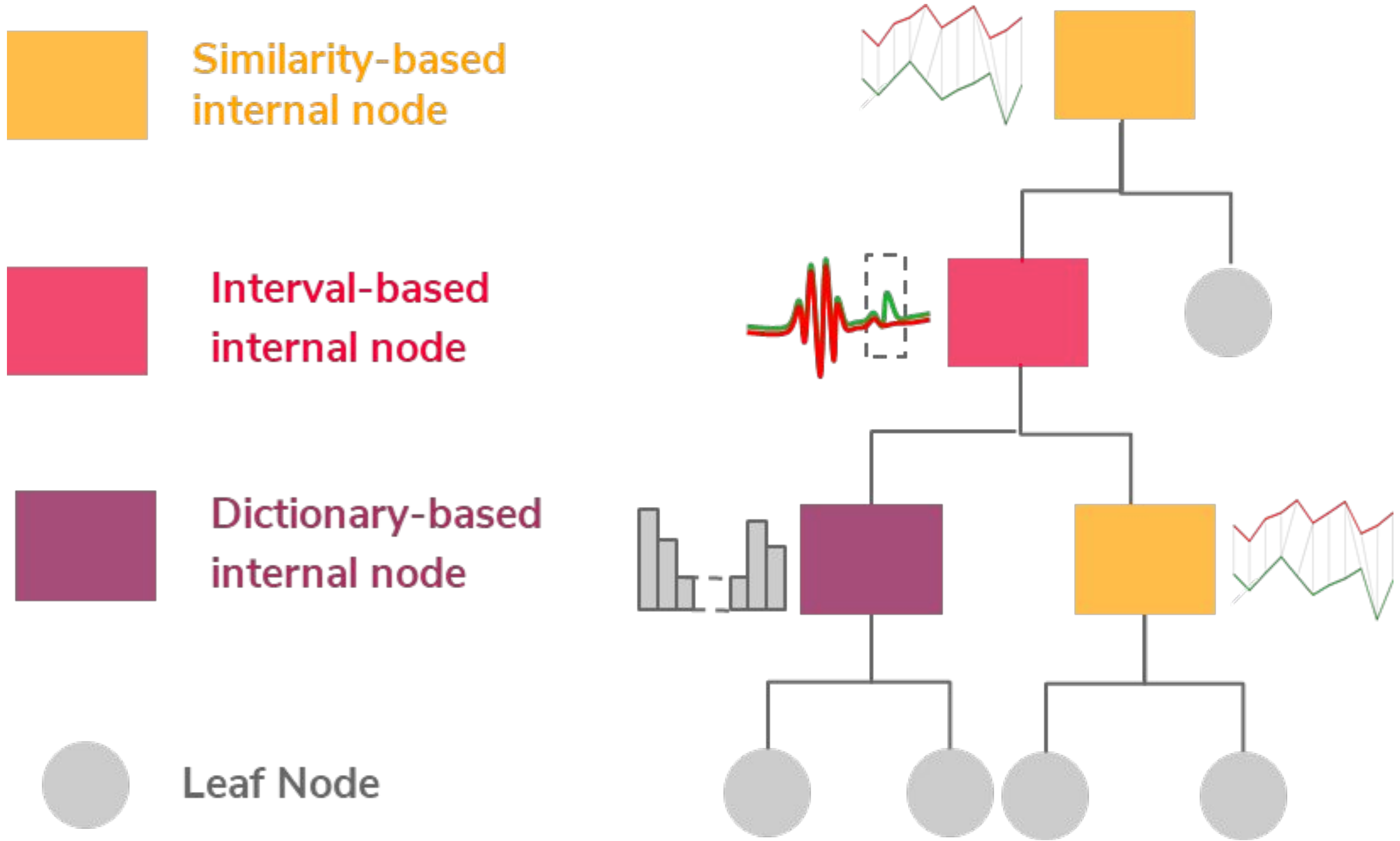
Child 2

Candidate split 3

- Candidates selected at random from all three strategies
- Selection using Gini Index

Time Series Combination of Heterogeneous Integrated Embeddings Forest (TS-CHIEF)

- TS-CHIEF trees combine three splitting functions
- Candidate splits selected at random
- Final selection using Gini Index



Similarity-based internal node

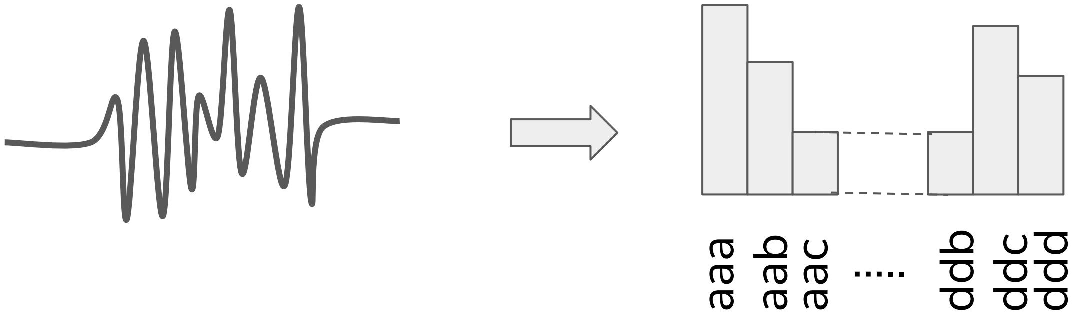
Interval-based internal node

Dictionary-based internal node

Leaf Node

TS-CHIEF : Dictionary-based splitter

- Precomputes a pool of BOSS transformations at forest level



- At node select a random transformation
- At node selects reference histograms per class (exemplars)
- Uses histogram similarity measure
- Partitions the data based on the proximity to reference histograms
- Original BOSS: Uses cross validation
- TS-CHIEF: Uses random transformations

Noname manuscript No.
(will be inserted by the editor)

The BOSS is concerned with time series classification in the presence of noise

Patrick Schäfer

the date of receipt and acceptance should be inserted later

Abstract Similarity search is one of the most important and probably best studied methods for data mining. In the context of time series analysis it reaches its limits when it comes to mining raw datasets. The raw time series data may be recorded at variable lengths, be noisy, or are composed of repetitive substructures. These build a foundation for state of the art search algorithms. However, noise has been paid surprisingly little attention to and is assumed to be filtered as part of a preprocessing step carried out by a human. Our Bag-of-SFA-Symbols (BOSS) model combines the extraction of substructures with the tolerance to extraneous and erroneous data using a noise reducing representation of the time series. We show that our BOSS ensemble classifier improves the best published classification accuracies in diverse application areas and on the official UCR classification benchmark datasets by a large margin.

Keywords Time Series · Classification · Similarity · Noise · Fourier Transform

1 Introduction

Time series are recorded from sensors and other input sources over time. Application domains include personalised medicine [22], human walking motions [28], anthropology [27], security [15], historical documents [27], astronomy [18], spectrographs [27], for example. While a human has an intuitive understanding of similarity, this task becomes very complex for a computer. It is non trivial to extract a statistical model from time series as these may be non-stationary, and show varying statistical properties with time. Data mining algorithms on the other hand, degenerate due to the high dimensionality of the time series and noise [11]. Existing techniques can be categorised as *shape-based* and *structure-based* [7]. Shape-based techniques use a *similarity measure* in combination with 1-nearest-neighbor search. These are competitive on pre-processed

Patrick Schäfer
Zuse Institute Berlin
Takustr. 7
14195 Berlin
Tel.: +49-30-84185-168
Fax: +49-30-84185-311
E-mail: patrick.schaefer@zib.de

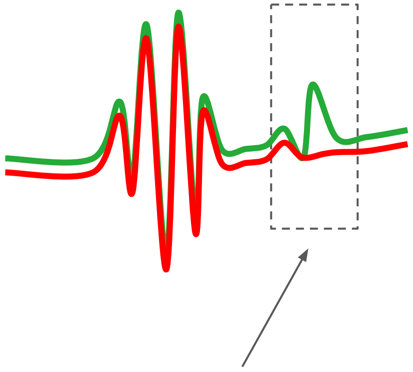
P. Schäfer (2015), The BOSS is concerned with time series classification in the presence of noise, *Data Mining and Knowledge Discovery*, Vol 29, Num 6, pp 1505–1530.

TS-CHIEF : Interval-based splitter

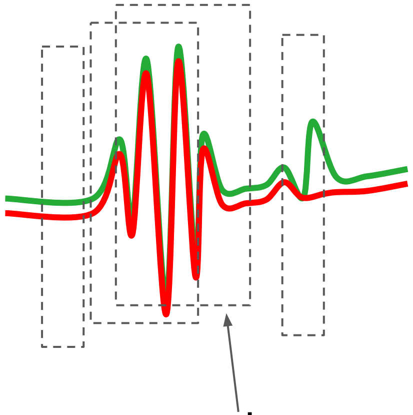
- Select random intervals and transforms
  - time (ACF, PACF, AR) and frequency (PS)
- Attribute-value split similar to classic decision tree

At tree level: RISE

At node level: TS-CHIEF

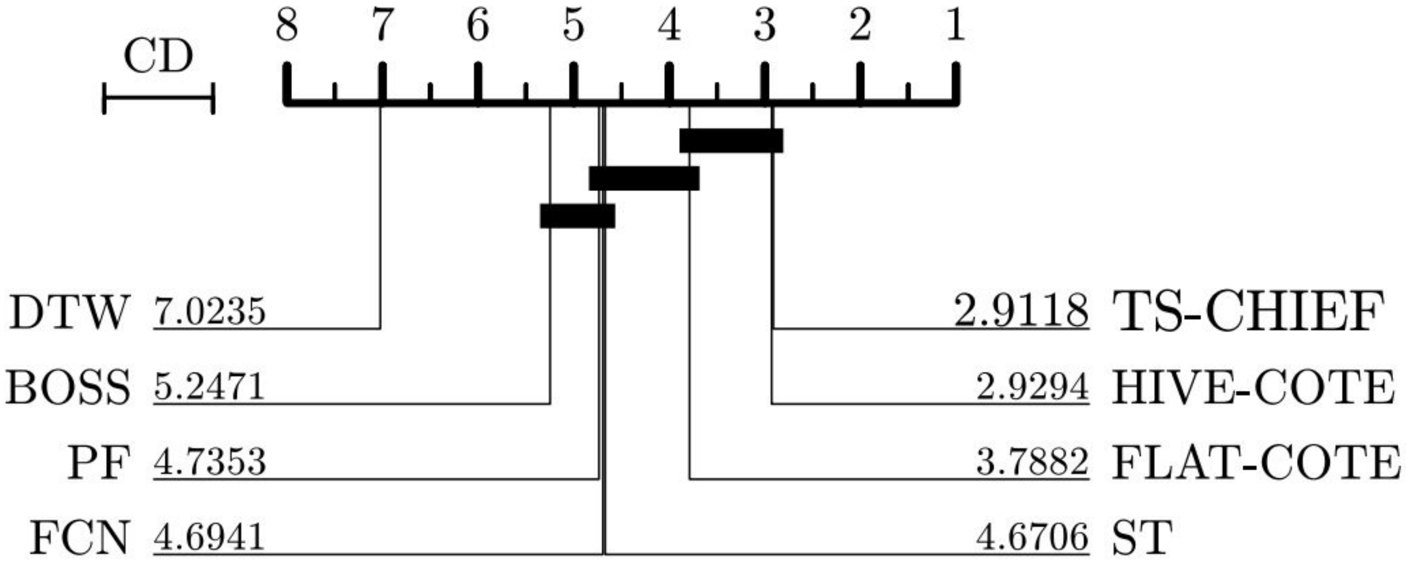


Intervals selected on global discrimination ability

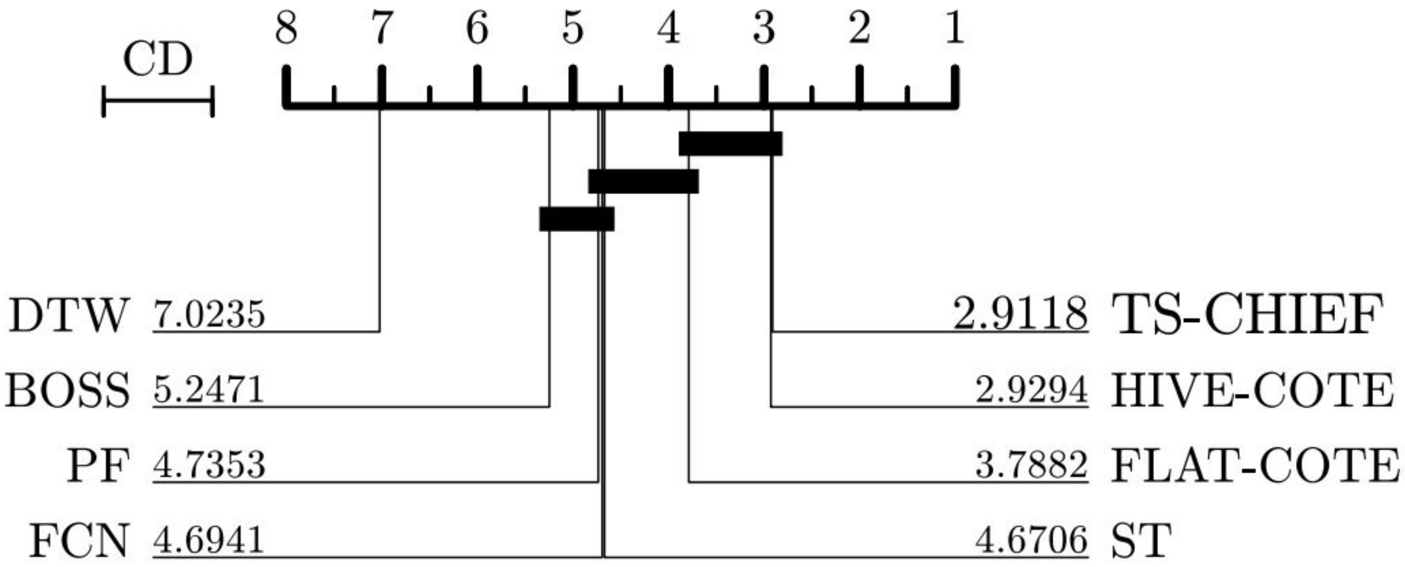


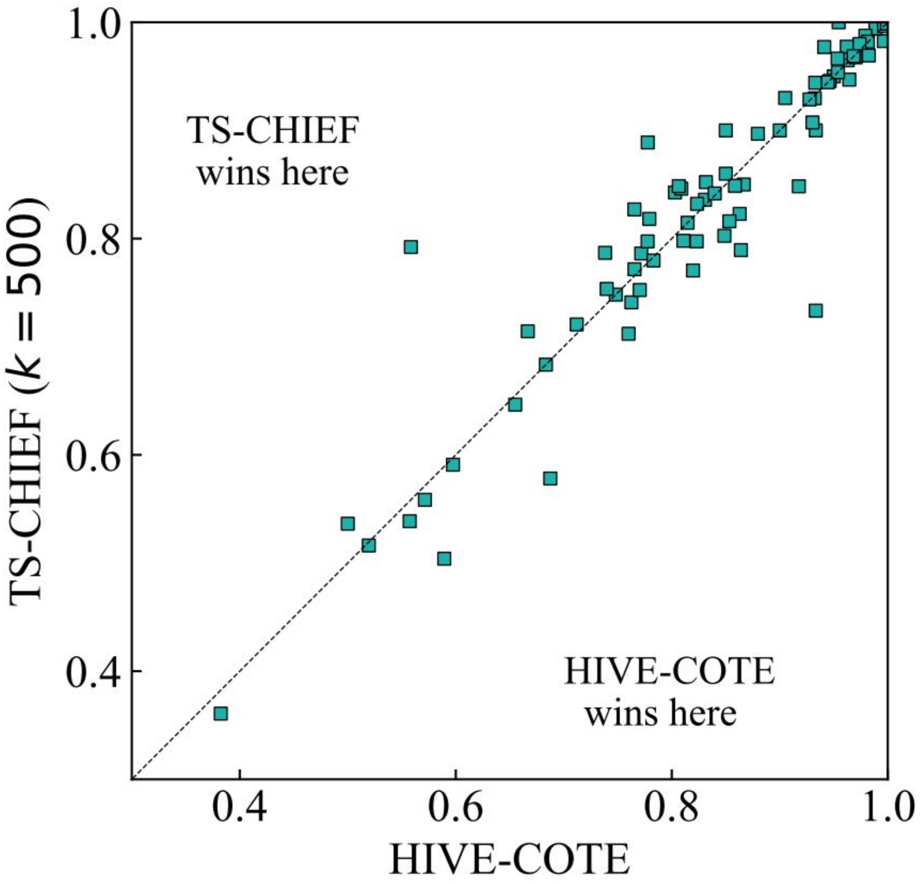
random intervals

Accuracy on 85 UCR datasets



Accuracy on 85 UCR datasets

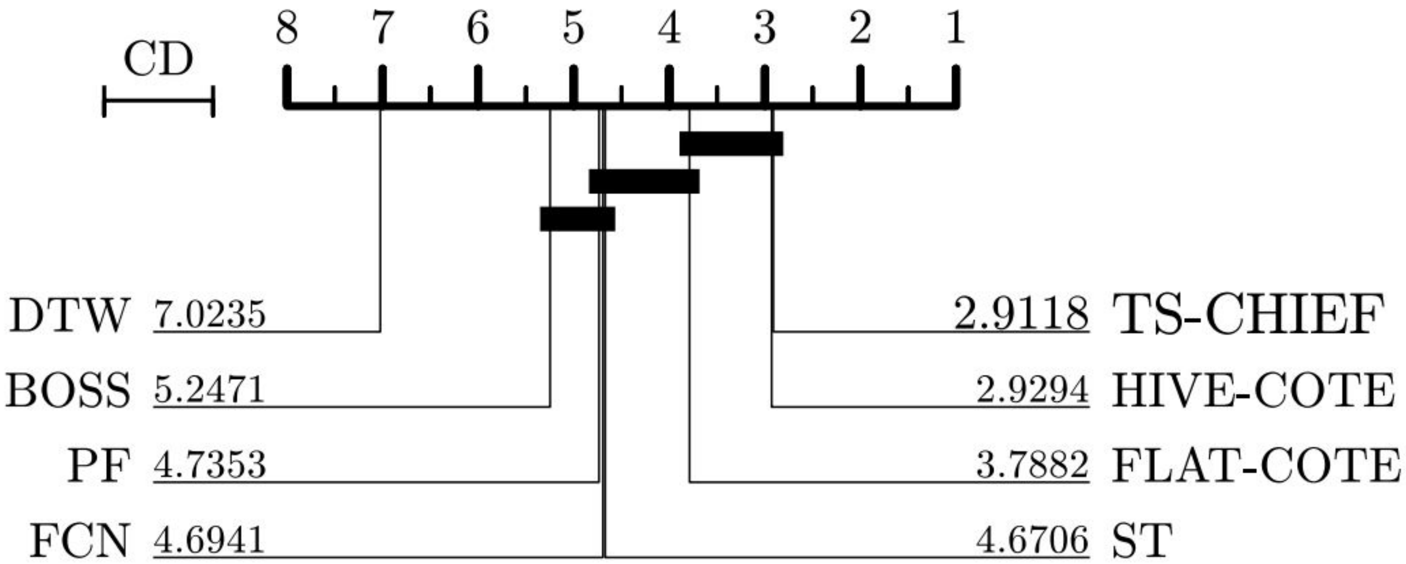


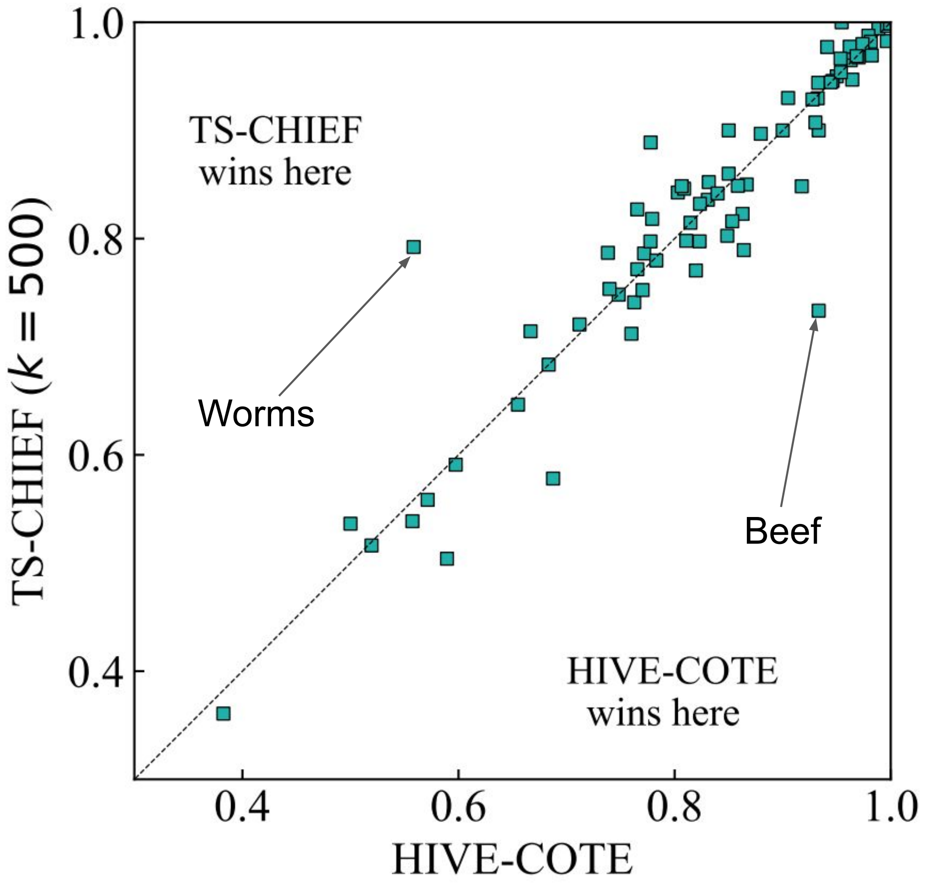


TS-CHIEF vs HIVE-COTE

41 wins
35 loss
9 ties
p=0.42 (Wilcoxon’s test)

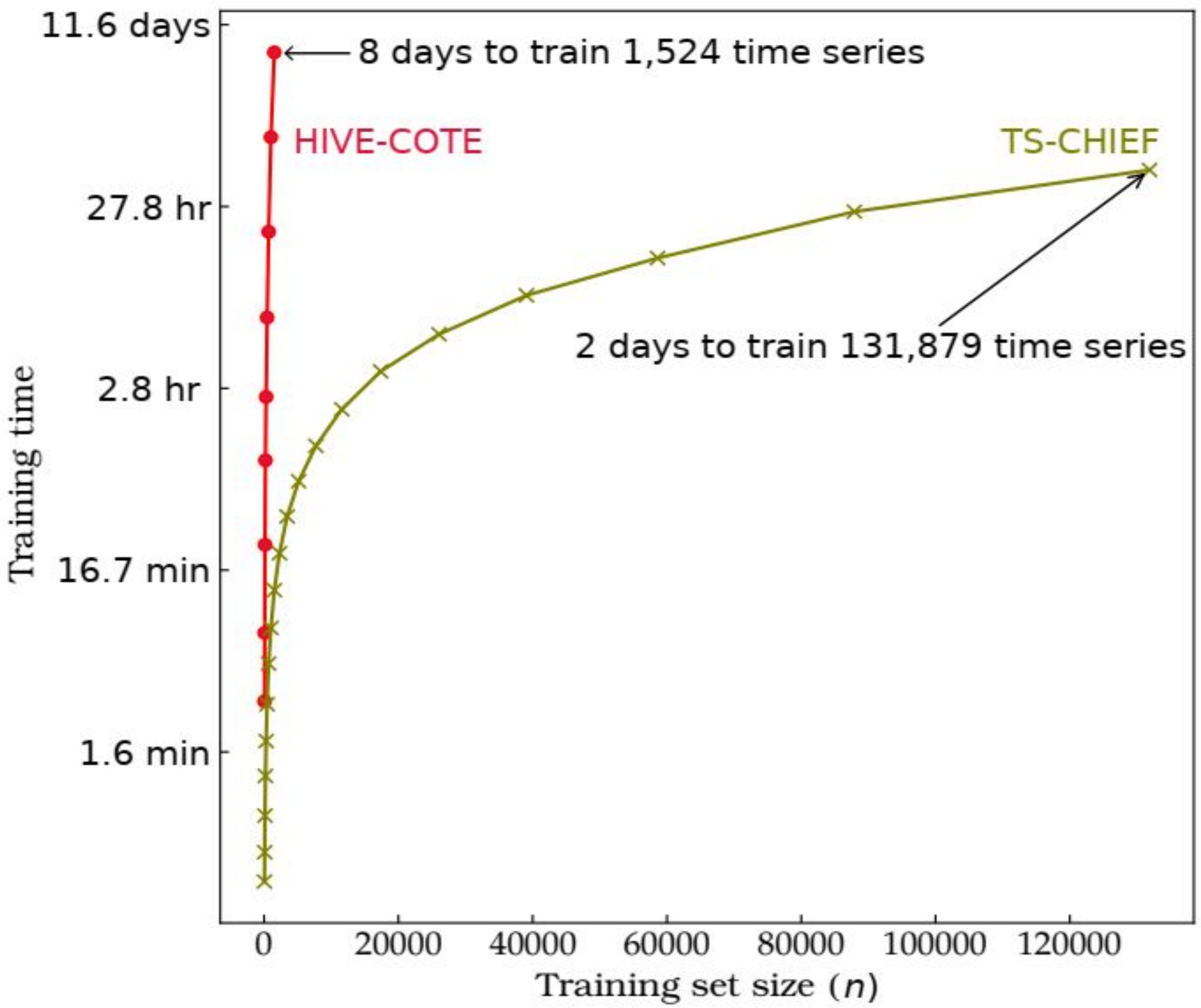
Accuracy on 85 UCR datasets





TS-CHIEF vs HIVE-COTE

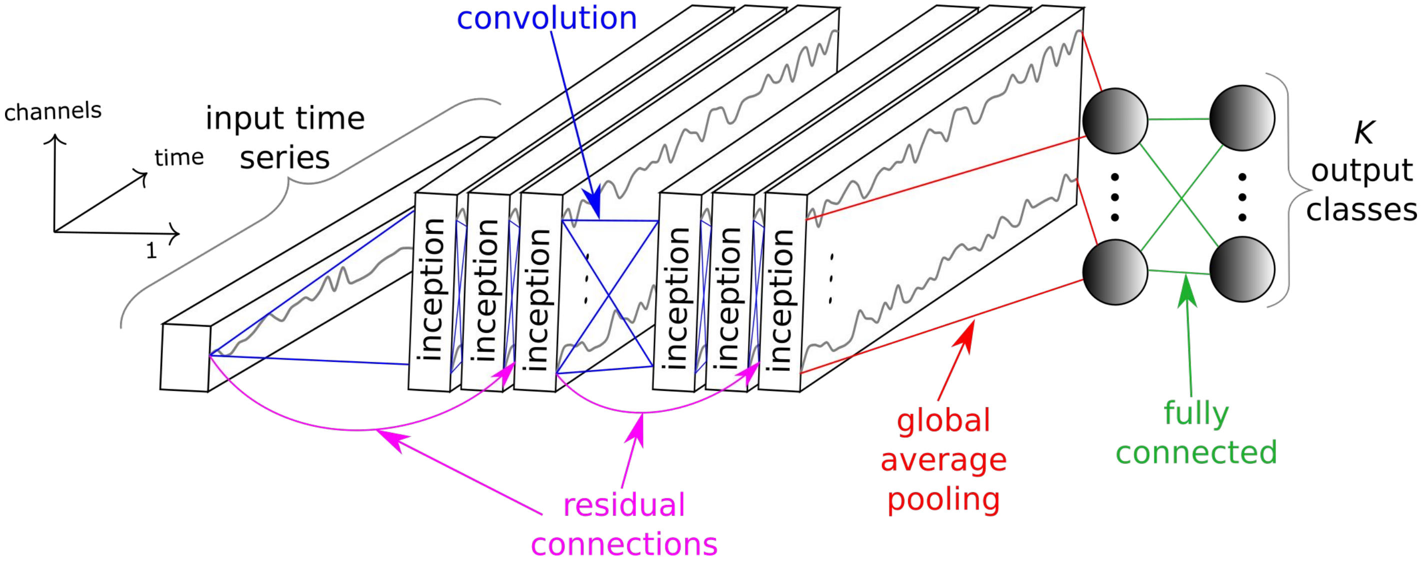
41 wins
35 loss
9 ties
p=0.42 (Wilcoxon's test)



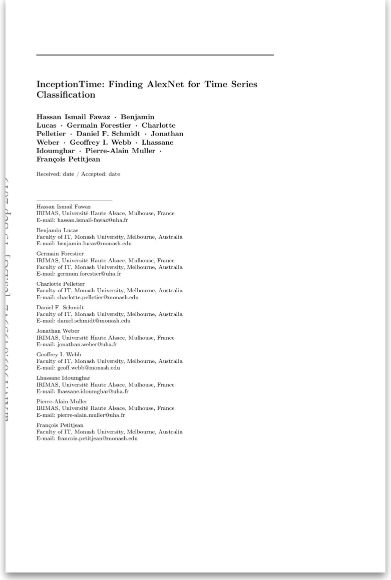
Training time vs training size

Training Size	HIVE-COTE	TS-CHIEF
1,500 time series	8 days	13 min (900x faster)
130,000 time series	230 years (estimated)	2 days (46,000x faster)

Part 3: InceptionTime



2019 - arxiv



InceptionTime: Finding AlexNet for Time Series Classification

Hassan Ismail Fawaz · Benjamin Lucas · Germain Forestier · Charlotte Pelletier · Daniel F. Schmidt · Jonathan Weber · Geoffrey I. Webb · Lhassane Idoumghar · Pierre-Alain Muller · François Petitjean

Received: date / Accepted: date

Hassan Ismail Fawaz
IRIMAS, Université Haute Alsace, Mulhouse, France
E-mail: hassan.ismail-fawaz@uha.fr

Benjamin Lucas
Faculty of IT, Monash University, Melbourne, Australia
E-mail: benjamin.lucas@monash.edu

Germain Forestier
IRIMAS, Université Haute Alsace, Mulhouse, France
Faculty of IT, Monash University, Melbourne, Australia
E-mail: germain.forestier@uha.fr

Charlotte Pelletier
Faculty of IT, Monash University, Melbourne, Australia
E-mail: charlotte.pelletier@monash.edu

Daniel F. Schmidt
Faculty of IT, Monash University, Melbourne, Australia
E-mail: daniel.schmidt@monash.edu

Jonathan Weber
IRIMAS, Université Haute Alsace, Mulhouse, France
E-mail: jonathan.weber@uha.fr

Geoffrey I. Webb
Faculty of IT, Monash University, Melbourne, Australia
E-mail: geoff.webb@monash.edu

Lhassane Idoumghar
IRIMAS, Université Haute Alsace, Mulhouse, France
E-mail: lhassane.idoumghar@uha.fr

Pierre-Alain Muller
IRIMAS, Université Haute Alsace, Mulhouse, France
E-mail: pierre-alain.muller@uha.fr

François Petitjean
Faculty of IT, Monash University, Melbourne, Australia
E-mail: francois.petitjean@monash.edu

http://bit.ly/InceptionTime

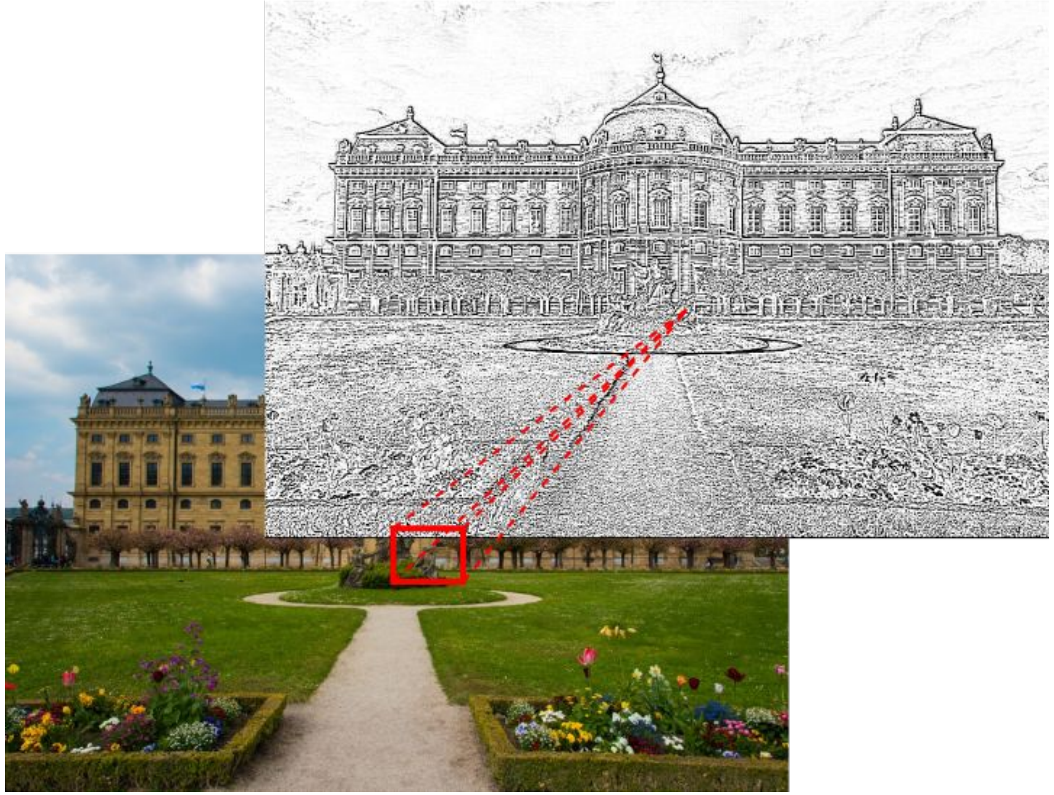
H. Ismail Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. Schmidt, J. Weber, G. Webb, L. Idoumghar, P-A. Muller, F. Petitjean (2019). InceptionTime: Finding AlexNet for Time Series Classification. *under review.* https://arxiv.org/abs/1909.04939

Deep Learning

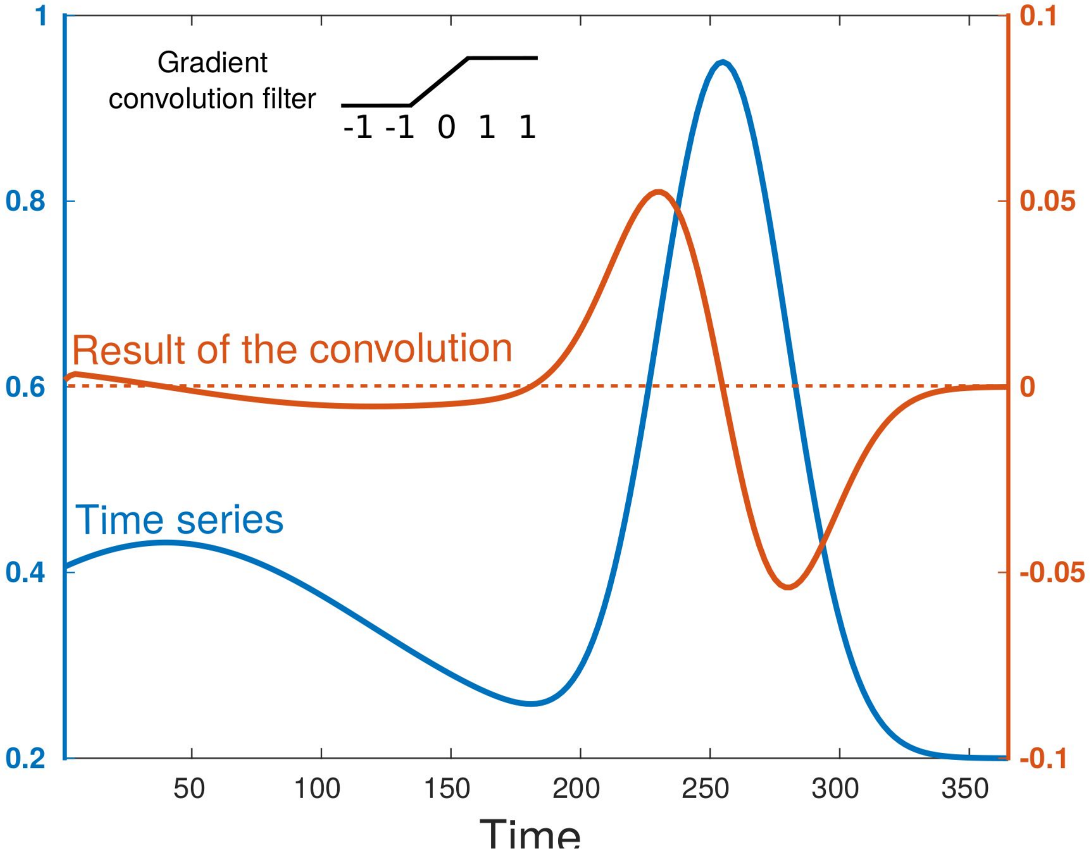
- Revolutionized the field of computer vision [1]
- Reached human level performance in image recognition tasks [2]
- Adopted by the Natural Language Processing (NLP) community [3]
- Improved state of the art speech recognition systems[4]

1. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems
2. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). Going deeper with convolutions. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition
3. Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. IEEE Computational intelligenCe magazine, 13(3), 55-75.
4. Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine

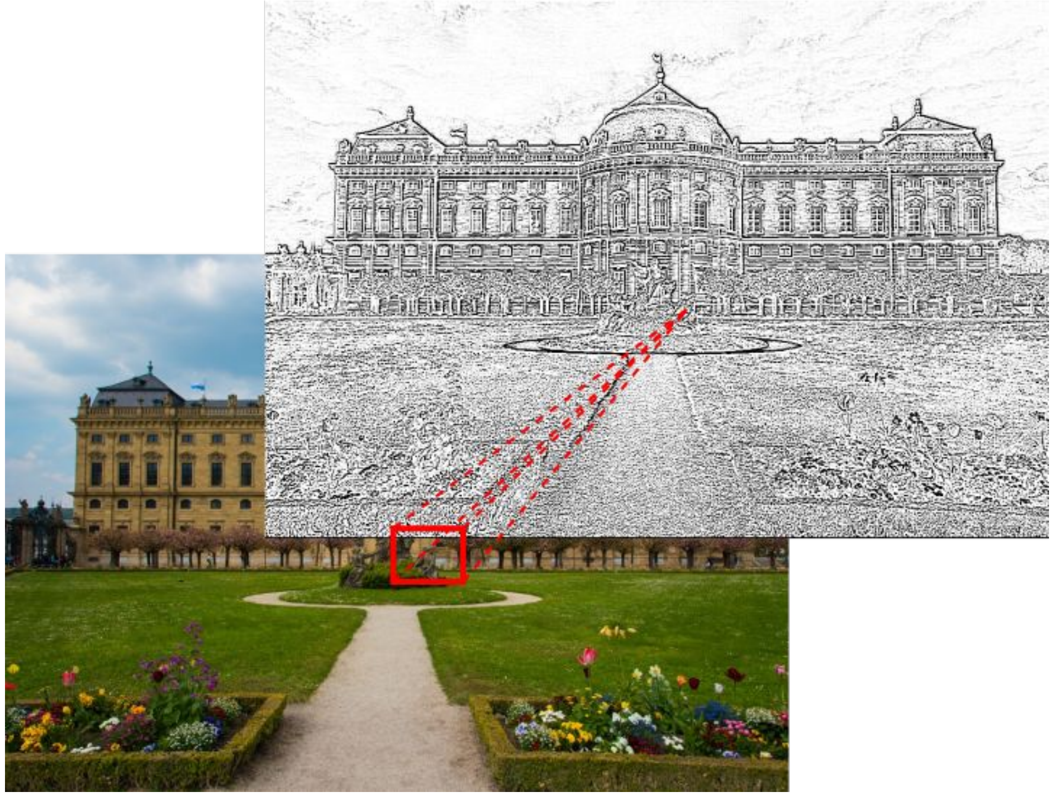
Convolution on images vs. time series



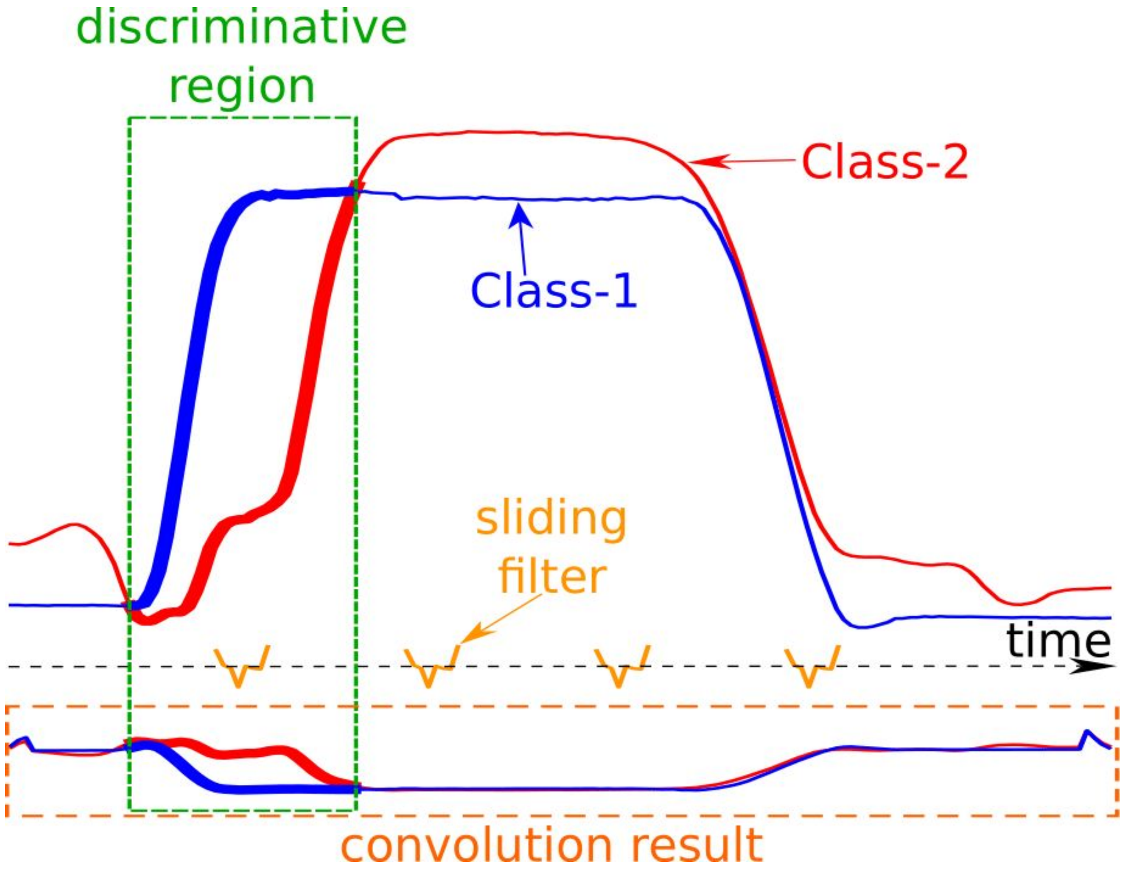
The result of a applying an edge detection convolution on an image



Convolution on images vs. time series

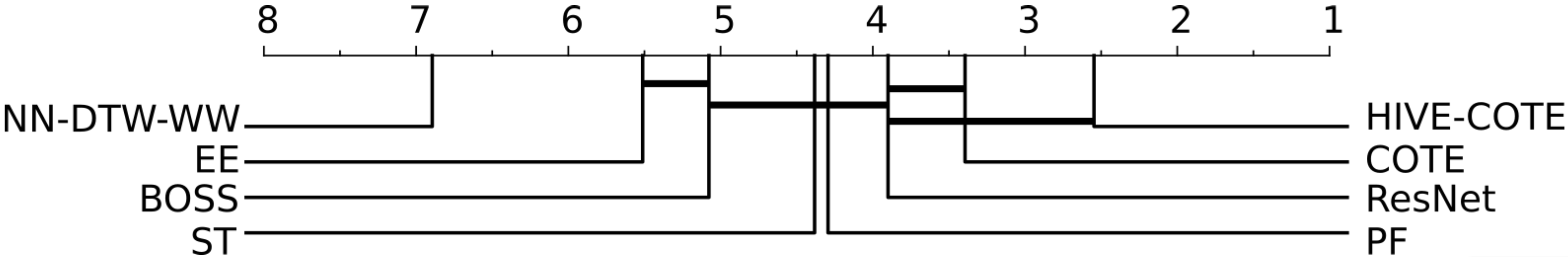


The result of a applying an edge detection convolution on an image



The result of applying a learned discriminative convolution on the GunPoint dataset

Deep learning for Time Series Classification



A critical difference diagram showing how ResNet still lacks behind the state of the art classifiers [1]

- Residual Network (ResNet) was originally proposed in [2]
- Currently is the state-of-the-art deep learning model for TSC [1]
- Designed to be a "baseline architecture" for TSC

Data Mining and Knowledge Discovery (2019) 33:917–963
https://doi.org/10.1007/s10618-019-00619-1

Deep learning for time series classification: a review

Hassan Ismail Fawaz[1] · Germain Forestier[1,2] · Jonathan Weber[1] ·
Lhassane Idoumghar[1] · Pierre-Alain Muller[1]

Received: 12 September 2018 / Accepted: 25 February 2019 / Published online: 2 March 2019

Abstract

Time Series Classification (TSC) is an important and challenging problem in data mining. With the increase of time series data availability, hundreds of TSC algorithms have been proposed. Among these methods, only a few have considered Deep Neural Networks (DNNs) to perform this task. This is surprising as deep learning has seen very successful applications in the last years. DNNs have indeed revolutionized the field of computer vision especially with the advent of novel deeper architectures such as Residual and Convolutional Neural Networks. Apart from images, sequential data such as text and audio can also be processed with DNNs to reach state-of-the-art performance for document classification and speech recognition. In this article, we study the current state-of-the-art performance of deep learning algorithms for TSC by presenting an empirical study of the most recent DNN architectures for TSC. We give an overview of the most successful deep learning applications in various time series domains under a unified taxonomy of DNNs for TSC. We also provide an open source deep learning framework to the TSC community where we implemented each of the compared approaches and evaluated them on a univariate TSC benchmark (the UCR/UEA archive) and 12 multivariate time series datasets. By training 8730 deep learning models on 97 time series datasets, we propose the most exhaustive study of DNNs for TSC to date.

Keywords Deep learning · Time series · Classification · Review

1 Introduction

During the last two decades, Time Series Classification (TSC) has been considered as one of the most challenging problems in data mining (Yang and Wu 2006; Esling

Responsible editor: Dr. Eamonn Keogh.

✉ Hassan Ismail Fawaz
hassan.ismail-fawaz@uha.fr

Extended author information available on the last page of the article

Springer

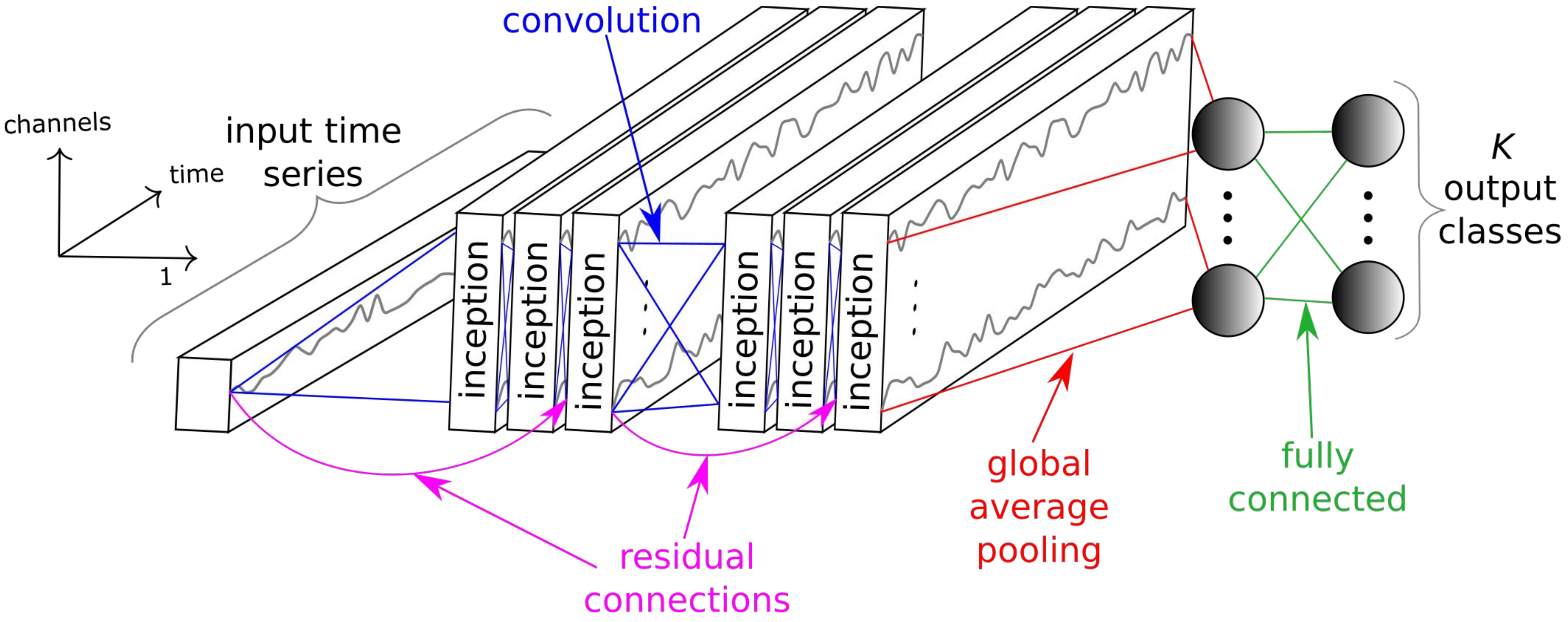
1. Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. A. (2019). Deep learning for time series classification: a review. Data Mining and Knowledge Discovery, 33(4), 917-963.
2. Wang, Z., Yan, W., & Oates, T. (2017, May). Time series classification from scratch with deep neural networks: A strong baseline. In IEEE I*nternational Joint Conference on Neural Networks.*

Inception

- Originally proposed by Google for image recognition problems [1]
- Further developed to reach state-of-the-art results on ImageNet [2]
- Main idea:
    - Apply convolutions of different resolutions to capture different patterns
    - Use a bottleneck layer in order to reduce the number of parameters
- For TSC, Inception had not been yet explored

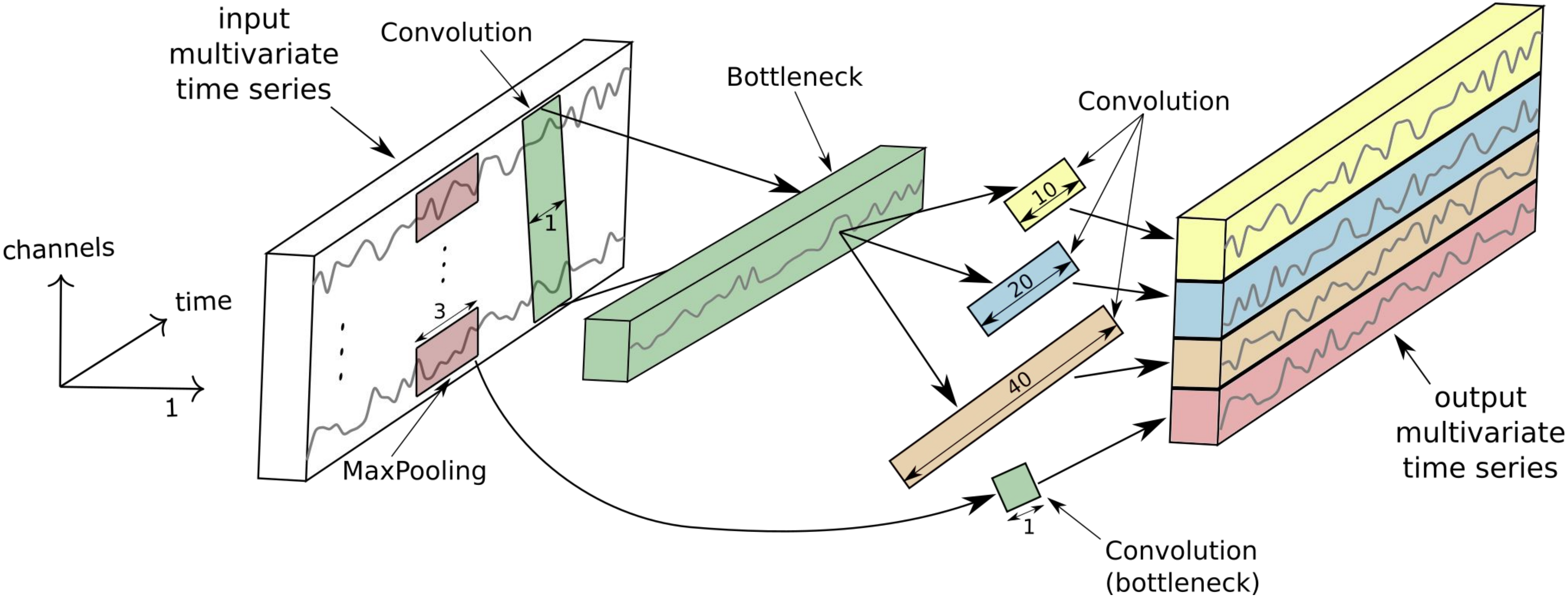
1. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).
2. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818-2826).

Our InceptionTime architecture for TSC



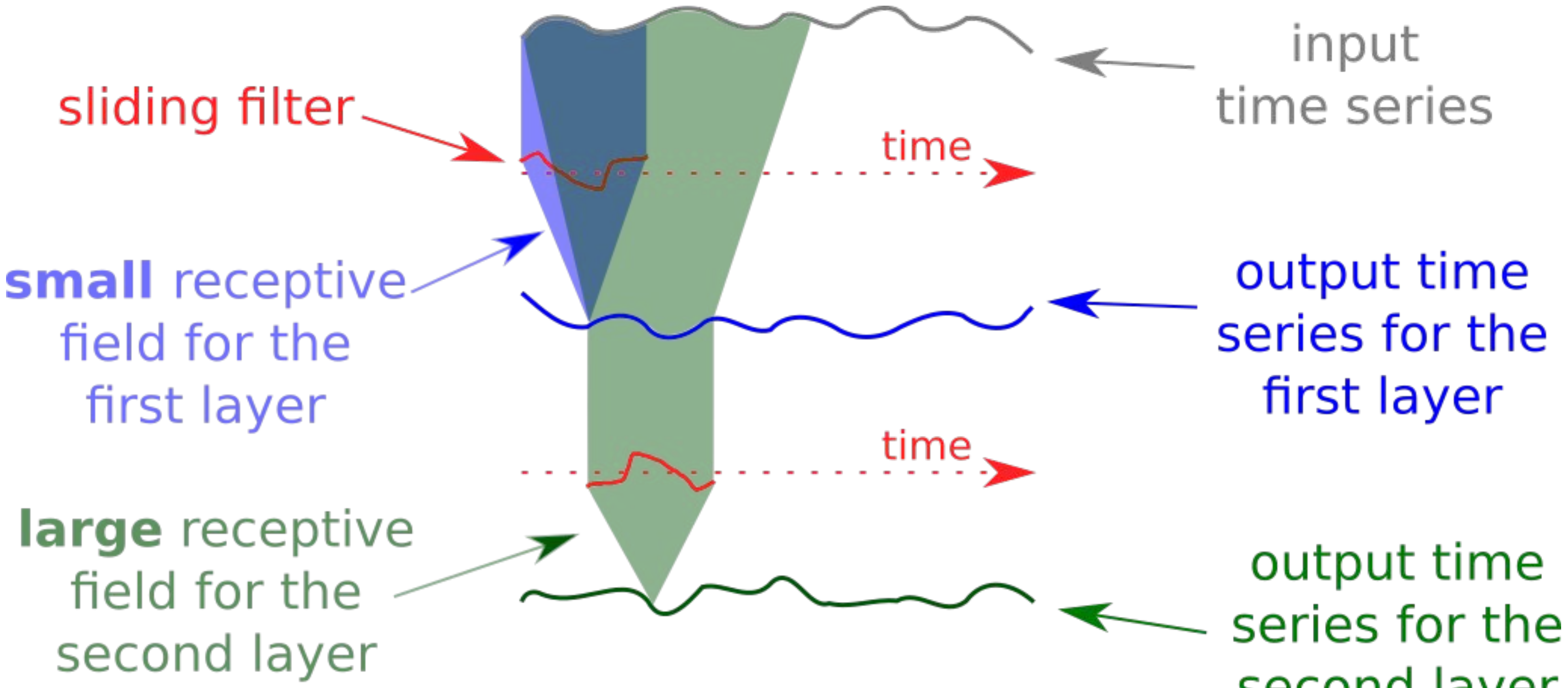
Inception network for time series classification

Inception **module** for time series classification



Inside our Inception module for time series classification

Receptive Field (RF) of a neural network



Receptive field illustration for a two layers CNN

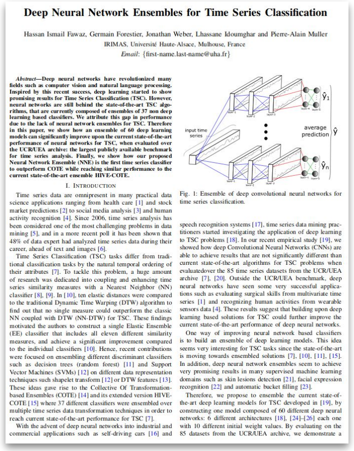
InceptionTime: an ensemble of 5 networks

$$\hat{y}_{i,c} = \frac{1}{n}\sum_{j=1}^{n}\sigma_c(x_i,\theta_j) \;\mid\; \forall c \in [1, C]$$

- Ensembling deep nets for TSC studied by Hassan in [1]
- Bias/variance tells us that this works because different initializations lead to very different networks (low covariance)

$$\mathrm{MSE}(H) = \overline{bias}(H)^2 + \frac{1}{|H|}\overline{variance}(H) + \left(1 - \frac{1}{|H|}\right)\overline{covariance}(H)$$

1. Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. (2019). Deep neural network ensembles for time series classification. IEEE International Joint Conference on Neural Networks.



Deep Neural Network Ensembles for Time Series Classification

Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar and Pierre-Alain Muller
IRIMAS, Université Haute-Alsace, Mulhouse, France
Email: {first-name.last-name@uha.fr}

***Abstract*—Deep neural networks have revolutionized many fields such as computer vision and natural language processing. Inspired by this recent success, deep learning started to show promising results for Time Series Classification (TSC). However, neural networks are still behind the state-of-the-art TSC algorithms, that are currently composed of ensembles of 37 non deep learning based classifiers. We attribute this gap in performance due to the lack of neural network ensembles for TSC. Therefore in this paper, we show how an ensemble of 60 deep learning models can significantly improve upon the current state-of-the-art performance of neural networks for TSC, when evaluated over the UCR/UEA archive: the largest publicly available benchmark for time series analysis. Finally, we show how our proposed Neural Network Ensemble (NNE) is the first time series classifier to outperform COTE while reaching similar performance to the current state-of-the-art ensemble HIVE-COTE.**

I. Introduction

Time series data are omnipresent in many practical data science applications ranging from health care [1] and stock market predictions [2] to social media analysis [3] and human activity recognition [4]. Since 2006, time series analysis has been considered one of the most challenging problems in data mining [5], and in a more recent poll it has been shown that 48% of data expert had analyzed time series data during their career, ahead of text and images [6].

Time Series Classification (TSC) tasks differ from traditional classification tasks by the natural temporal ordering of their attributes [7]. To tackle this problem, a huge amount of research was dedicated into coupling and enhancing time series similarity measures with a Nearest Neighbor (NN) classifier [8], [9]. In [10], ten elastic distances were compared to the traditional Dynamic Time Warping (DTW) algorithm to find out that no single measure could outperform the classic NN coupled with DTW (NN-DTW) for TSC. These findings motivated the authors to construct a single Elastic Ensemble (EE) classifier that includes all eleven different similarity measures, and achieve a significant improvement compared to the individual classifiers [10]. Hence, recent contributions were focused on ensembling different discriminant classifiers such as decision trees (random forest) [11] and Support Vector Machines (SVMs) [12] on different data representation techniques such shapelet transform [12] or DTW features [13]. These ideas gave rise to the Collective Of Transformation-based Ensembles (COTE) [14] and its extended version HIVE-COTE [15] where 37 different classifiers were ensembled over multiple time series data transformation techniques in order to reach current state-of-the-art performance for TSC [7].

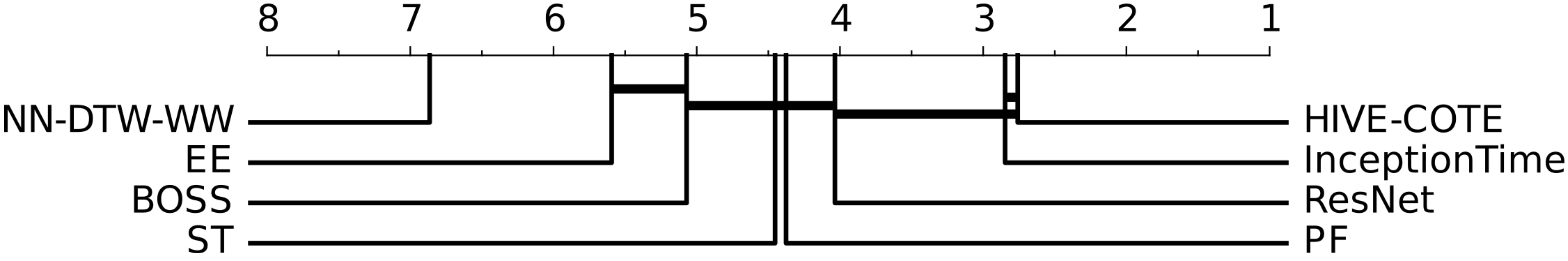
With the advent of deep neural networks into industrial and commercial applications such as self-driving cars [16] and speech recognition systems [17], time series data mining practitioners started investigating the application of deep learning to TSC problems [18]. In our recent empirical study [19], we showed how deep Convolutional Neural Networks (CNNs) are able to achieve results that are not significantly different than current state-of-the-art algorithms for TSC problems when evaluatedover the 85 time series datasets from the UCR/UEA archive [7], [20]. Outside the UCR/UEA benchmark, deep neural networks have seen some very successful applications such as evaluating surgical skills from multivariate time series [1] and recognizing human activities from wearable sensors data [4]. These results suggest that building upon deep learning based solutions for TSC could further improve the current state-of-the-art performance of deep neural networks.

One way of improving neural network based classifiers is to build an ensemble of deep learning models. This idea seems very interesting for TSC tasks since the state-of-the-art is moving towards ensembled solutions [7], [10], [11], [15]. In addition, deep neural network ensembles seem to achieve very promising results in many supervised machine learning domains such as skin lesions detection [21], facial expression recognition [22] and automatic bucket filling [23].

Therefore, we propose to ensemble the current state-of-the-art deep learning models for TSC developed in [19], by constructing one model composed of 60 different deep neural networks: 6 different architectures [18], [24]–[26] each one with 10 different initial weight values. By evaluating on the 85 datasets from the UCR/UEA archive, we demonstrate a

Fig. 1: Ensemble of deep convolutional neural networks for time series classification.

Accuracy results on the UCR archive



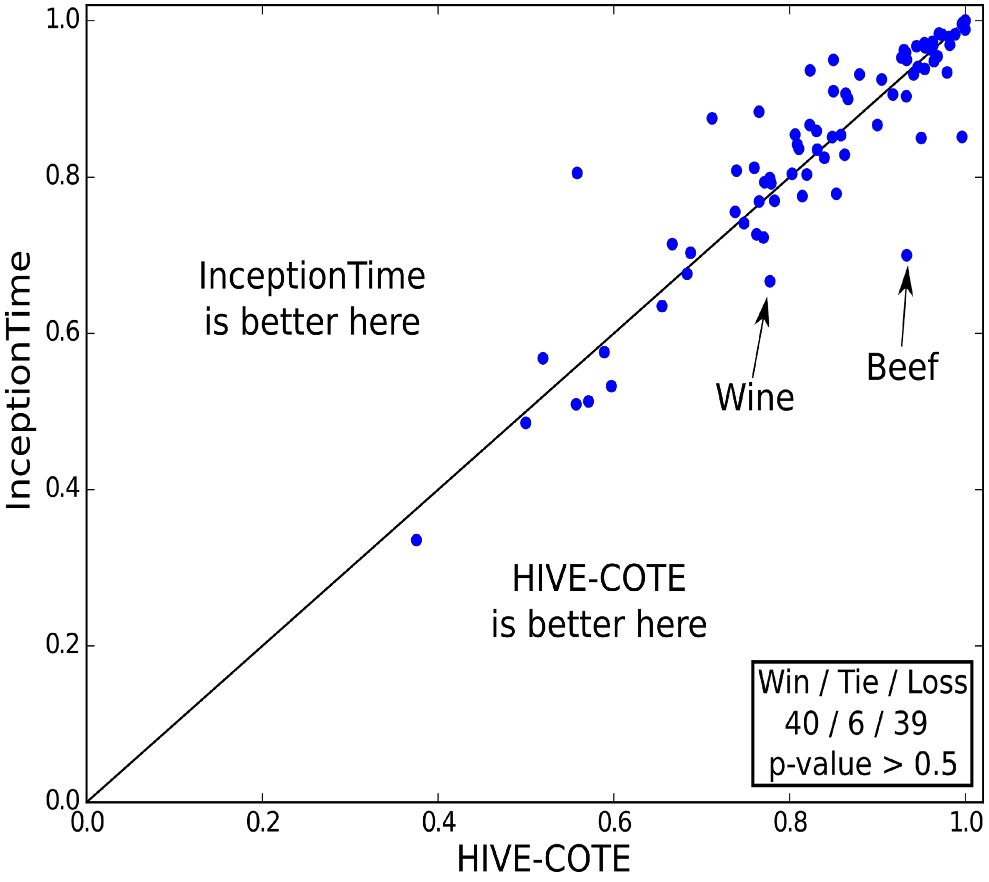
Critical difference diagram showing the performance of InceptionTime compared to the current state-of-the-art classifiers of time series data

- InceptionTime reaches very similar results to HIVE-COTE

Bagnall, A., Lines, J., Bostrom, A., Large, J., & Keogh, E. (2017). The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, *31*(3), 606-660.

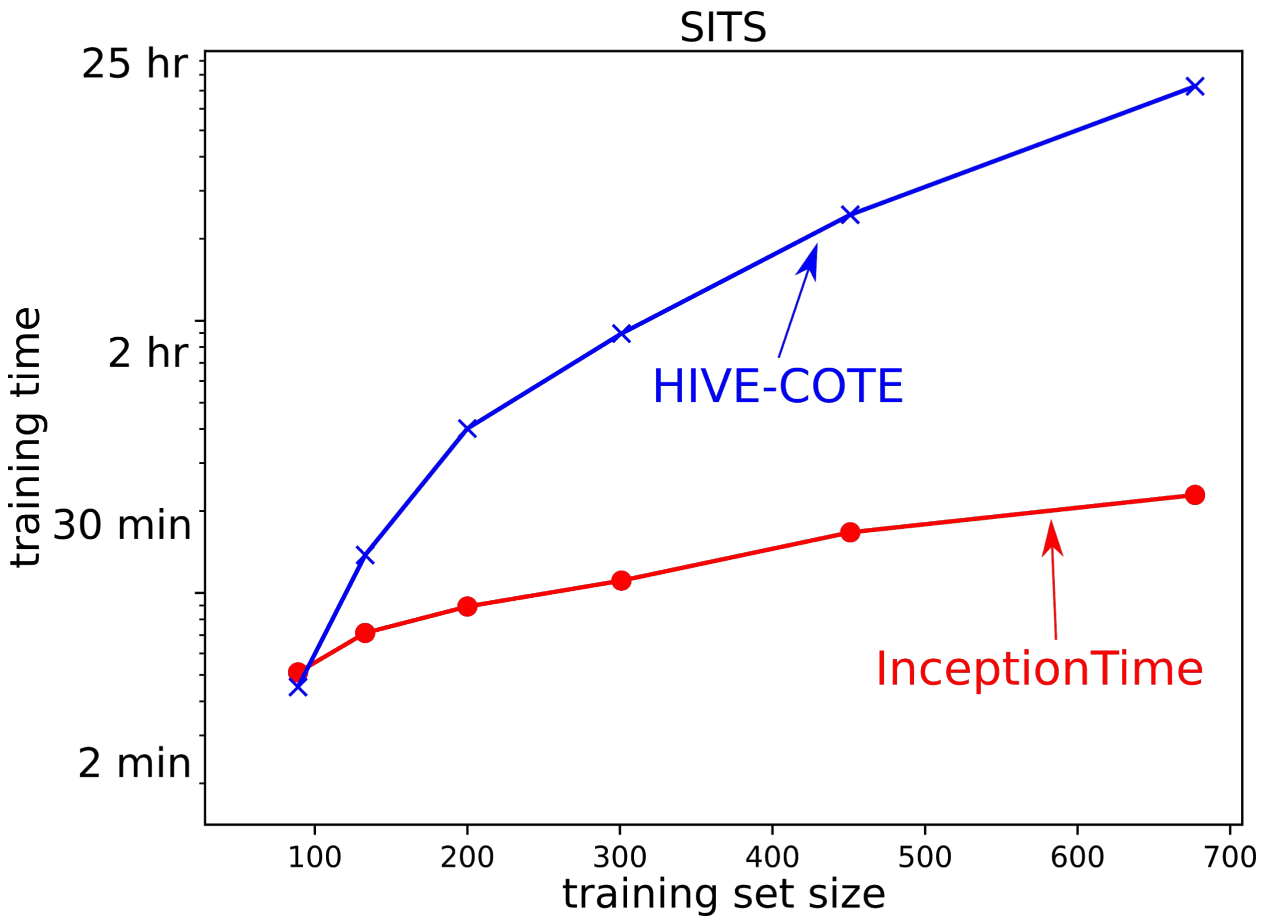
Accuracy plot: InceptionTime vs HIVE-COTE

- InceptionTime is slightly better than HIVE-COTE on average [1]
- Wine and Beef were shown to benefit from transfer learning [2]

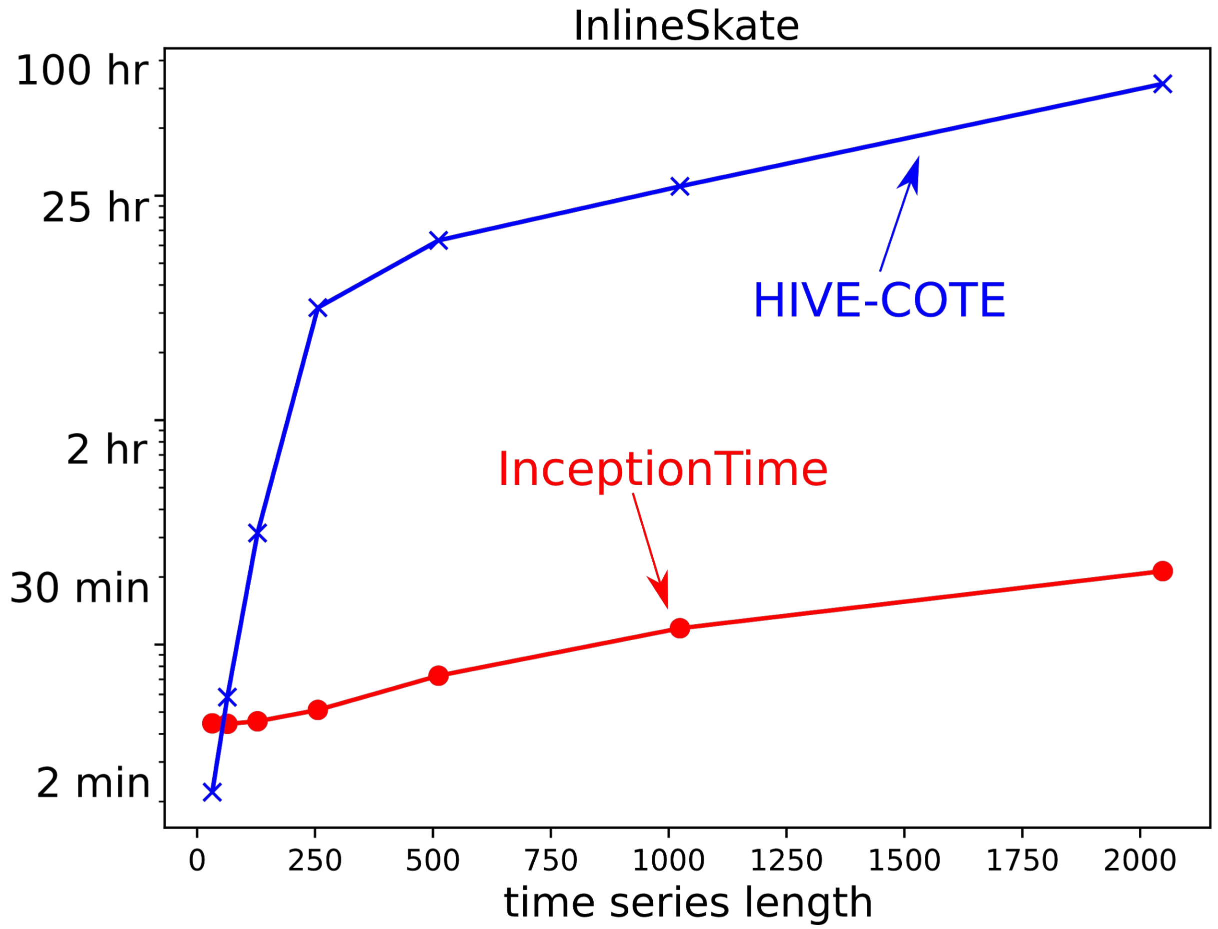


1. Lines, J., Taylor, S., & Bagnall, A. (2018). Time series classification with HIVE-COTE: The hierarchical vote collective of transformation-based ensembles. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *12*(5), 52.
2. Ismail Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. A. (2018). Transfer learning for time series classification. In IEEE International Conference on Big Data.

Training time comparison with HIVE-COTE

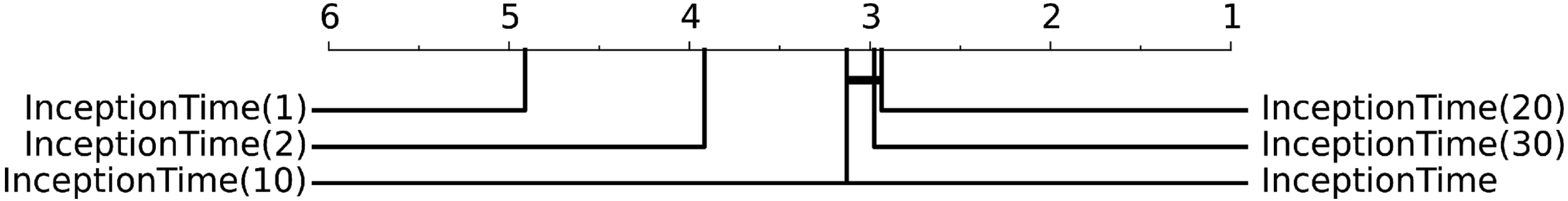


Training time as a function of the training set size for the SITS dataset



Training time as a function of the series length for the InlineSkate dataset

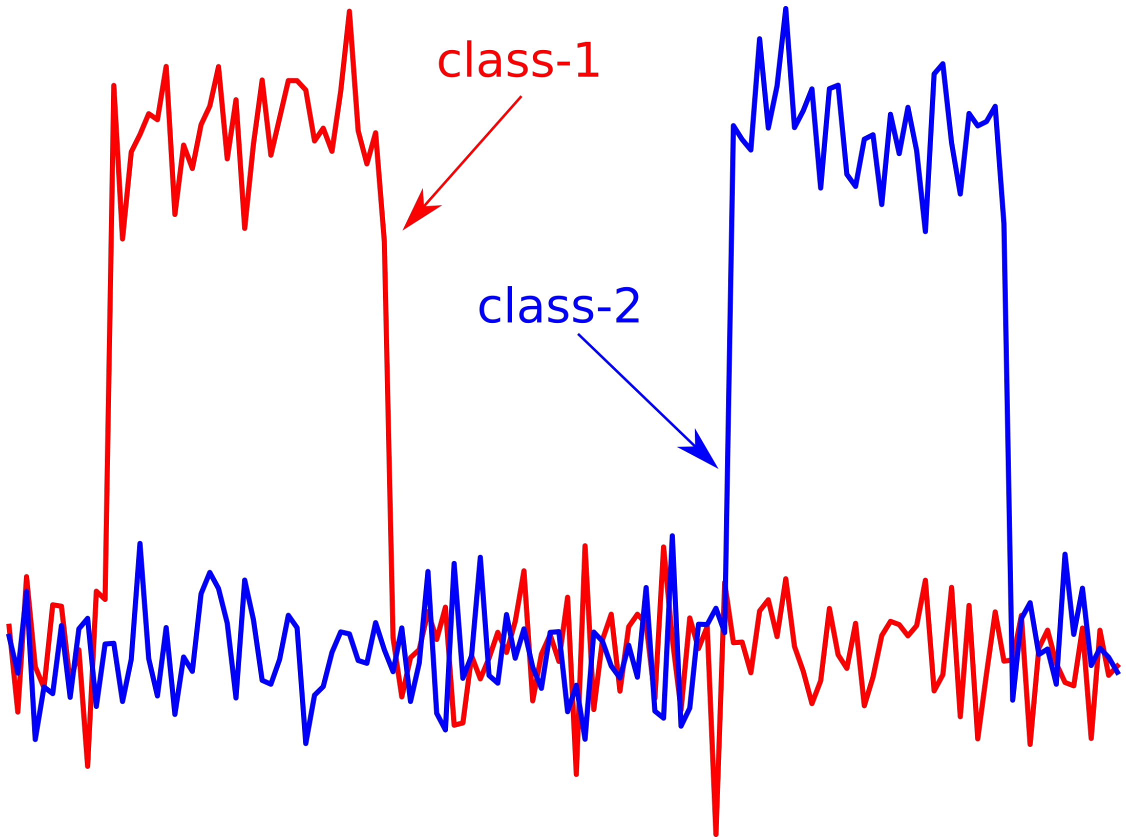
Studying the size of the ensemble



Critical difference diagram showing the effect of the number of individual classifiers in InceptionTime

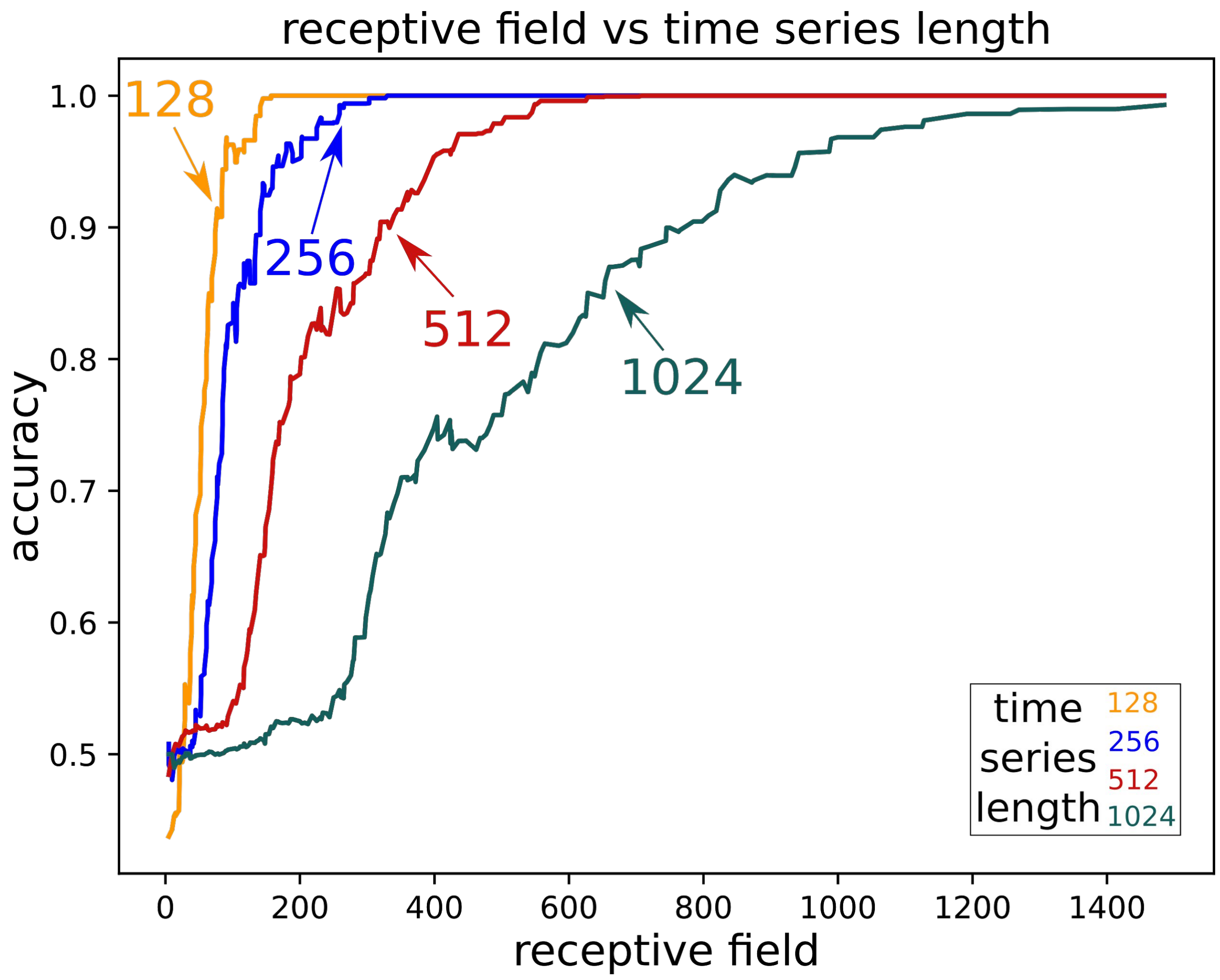
- InceptionTime(*x*) denotes an ensemble of *x* Inception networks
- InceptionTime is equivalent to InceptionTime(5)
- There is no significant improvement for x≥5
  - ○ Again this is due to covariance that start hurting us from 5 elements
  - ○ Therefore we decided to stick with InceptionTime(5)

Hyperparameter study: synthetic dataset



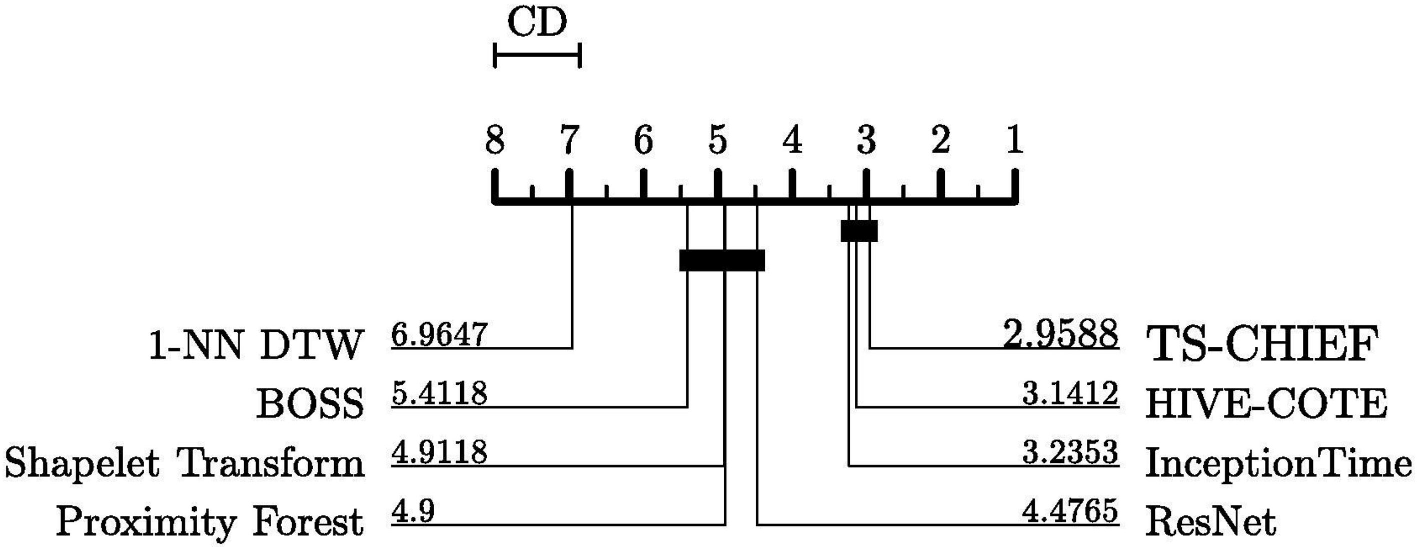
Example of a synthetic binary time series classification problem

Hyperparameter study: Receptive Field (RF)

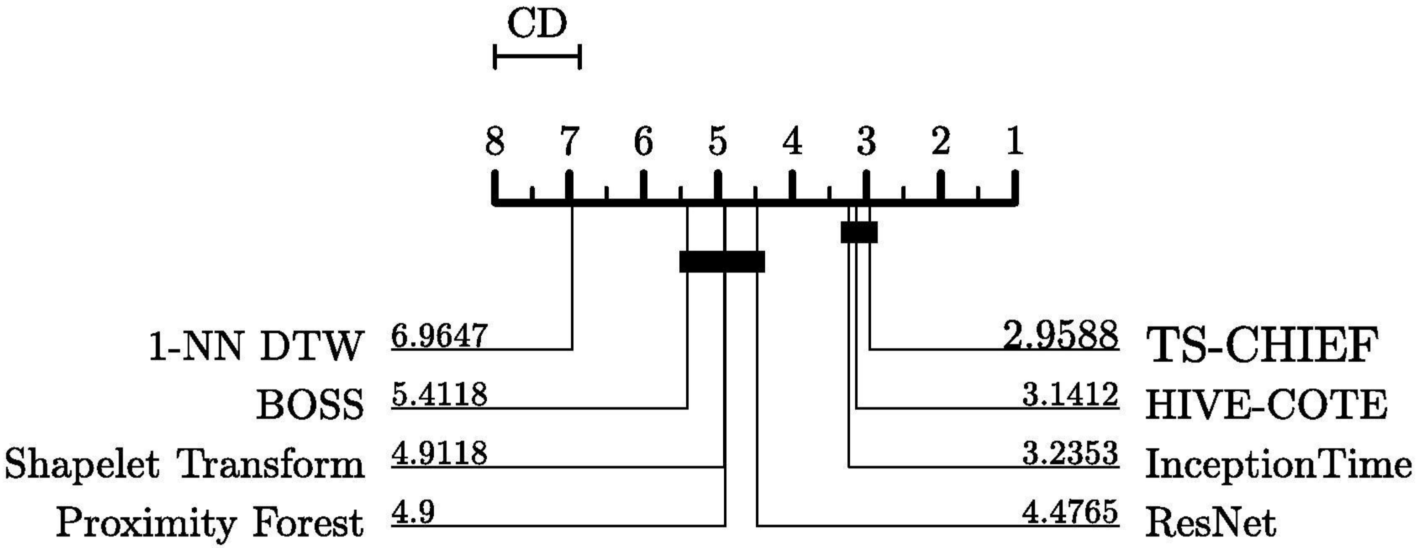


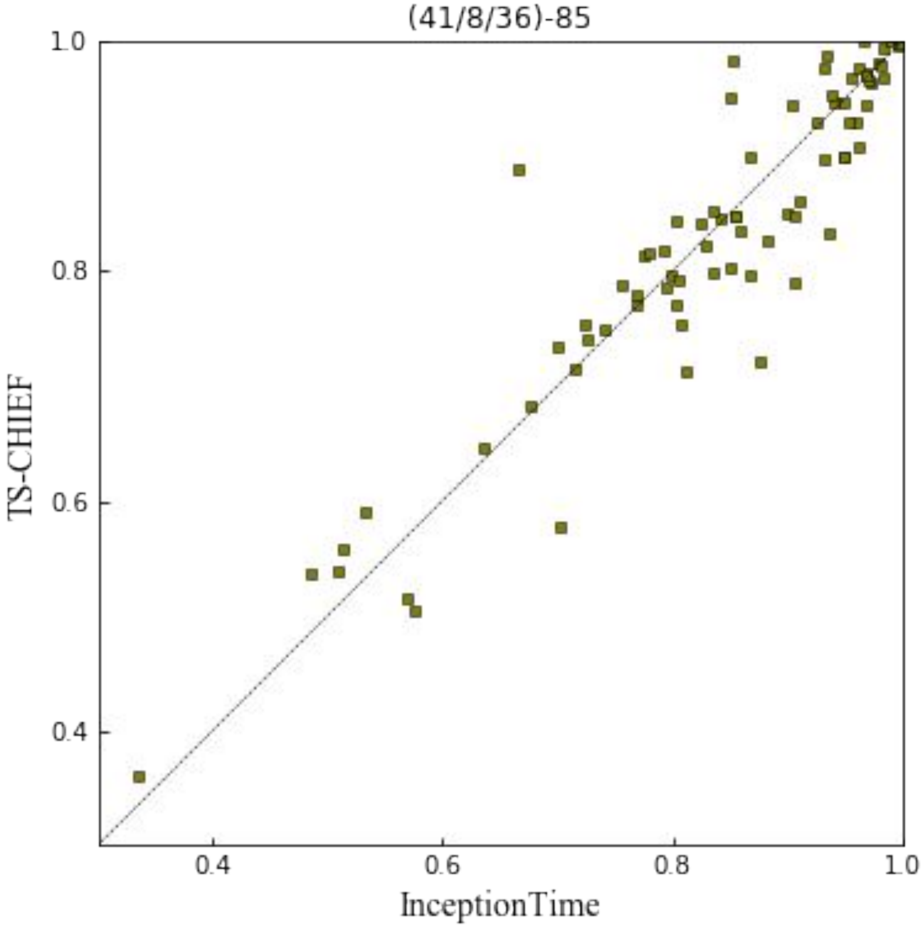
A larger receptive field is needed to classify very long time series

TS-CHIEF vs InceptionTime



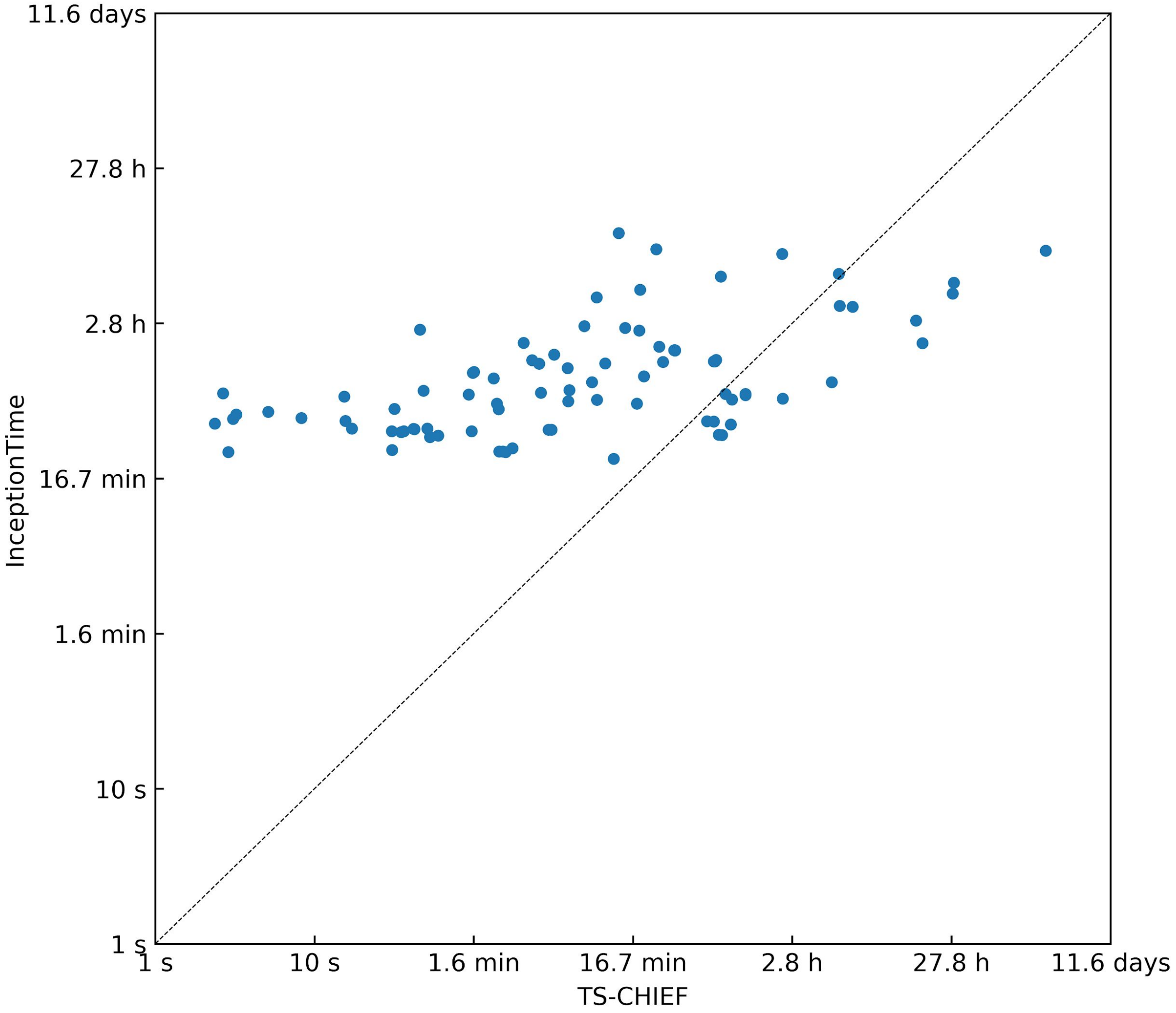
TS-CHIEF vs InceptionTime



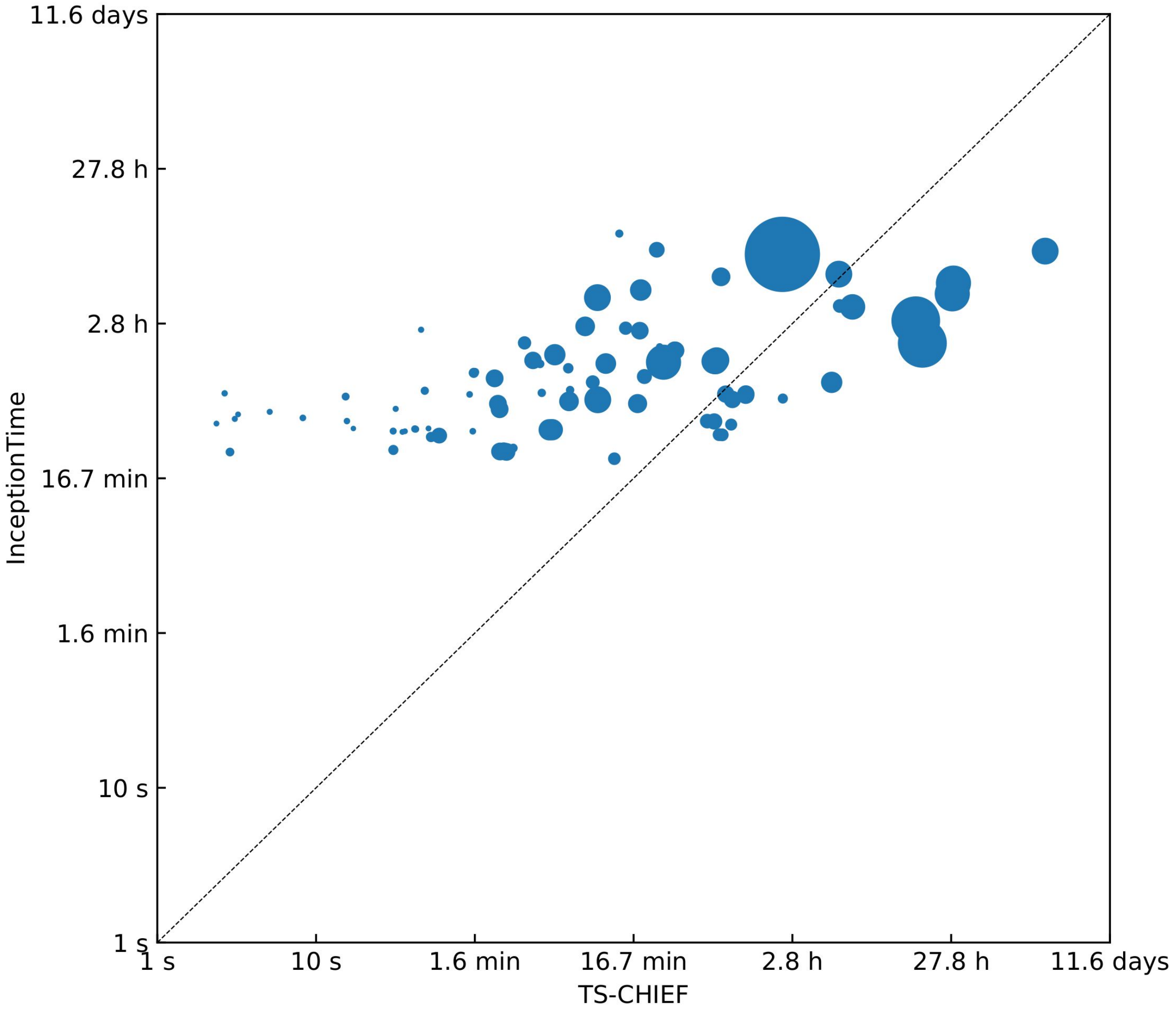


- TS-CHIEF wins on average in terms of WDL
- But, lots of big wins with InceptionTime

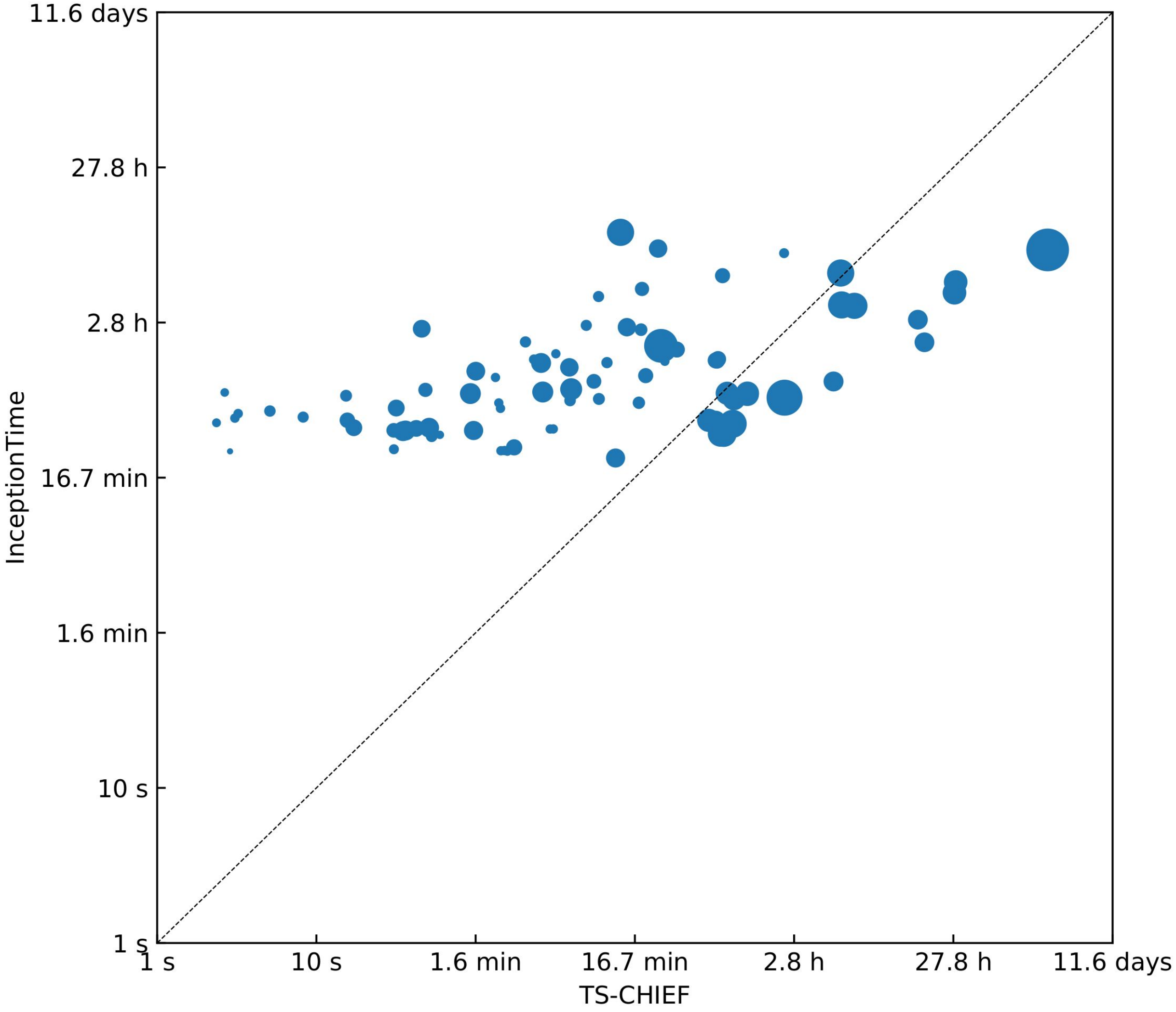
TS-CHIEF vs InceptionTime - training time



TS-CHIEF vs InceptionTime - training time



TS-CHIEF vs InceptionTime - training time



Conclusions

- Ensemble techniques have revolutionised time series classification with Tony and Jason's group giving us a beacon for research
- TS-Chief combines the efficiencies of tree-based divide-and-conquer with random split selection and the effectiveness of decades' worth of specialised time series techniques
- InceptionTime brings the power and efficiency of deep learning
- Both make state-of-the-art accuracy computationally feasible for large learning tasks
- We believe in reproducible research:
  - Proximity Forest ⟶ https://github.com/fpetitjean/ProximityForest
  - TS-CHIEF ⟶ https://github.com/dotnet54/TS-CHIEF
  - InceptionTime ⟶ https://github.com/hfawaz/InceptionTime

MONASH University

JOIN US IN MELBOURNE!

- 2.5-year postdoc in ML - http://bit.ly/JobsFrancois
- 3 year postdoc+dev in time series

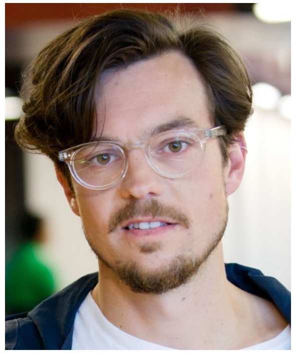
3,600 € per month after-tax

PhD positions available

Send me an email if interested

→ francois.petitjean@monash.edu

MONASH University



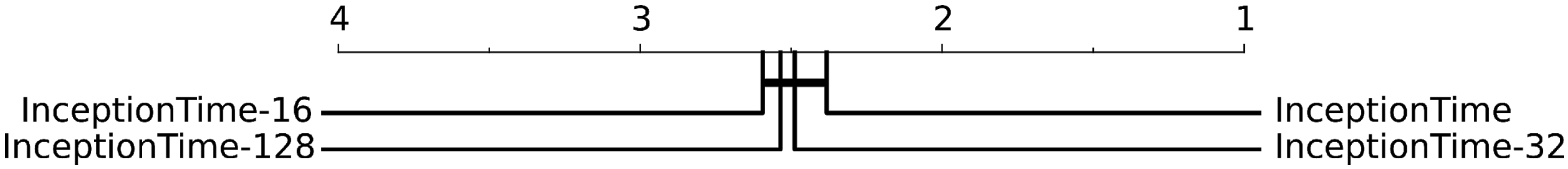
Thank you!

http://francois-petitjean.com

GROUP OF EIGHT AUSTRALIA

Additional slides

Hyperparameter study: batch size

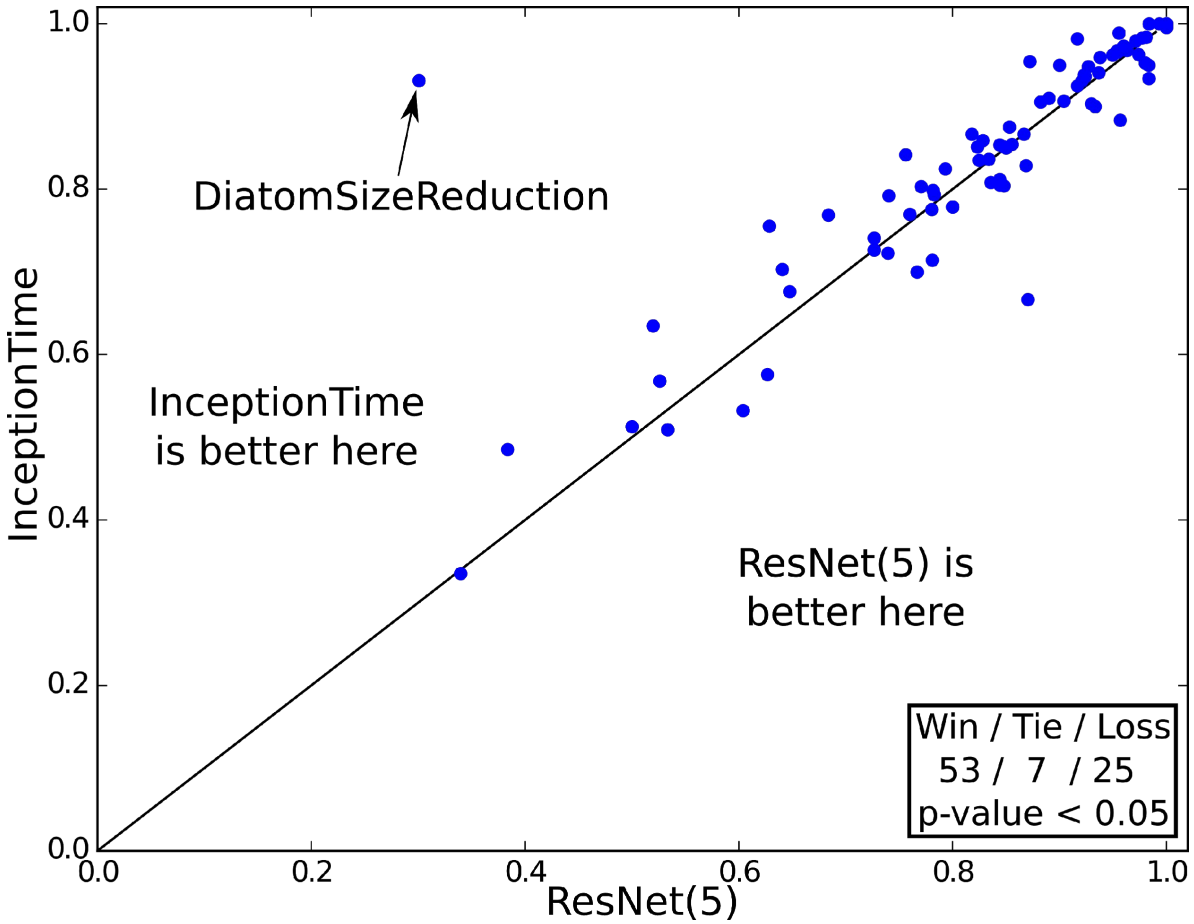


Critical difference diagram showing the effect of the batch size hyperparameter value over InceptionTime's average rank

- InceptionTime-*x* denotes InceptionTime with a batch size equal to *x*
- InceptionTime is equivalent to InceptionTime-64 (default value)
- There is no significant difference between the different models
- A value equal to 64 shows a small non-significant superiority
- We therefore chose to stick with a batch size equal to 64

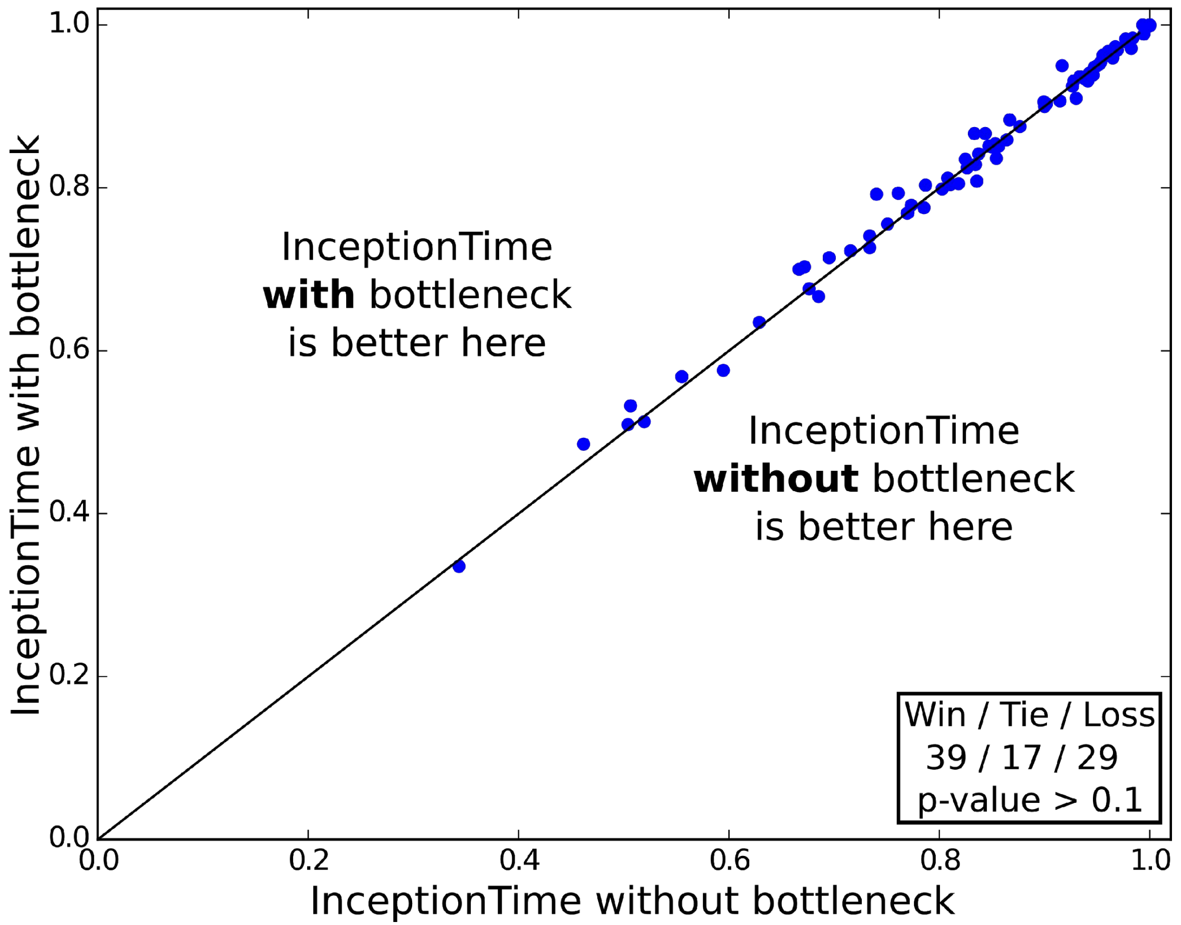
Accuracy plot: InceptionTime vs ResNet(5)

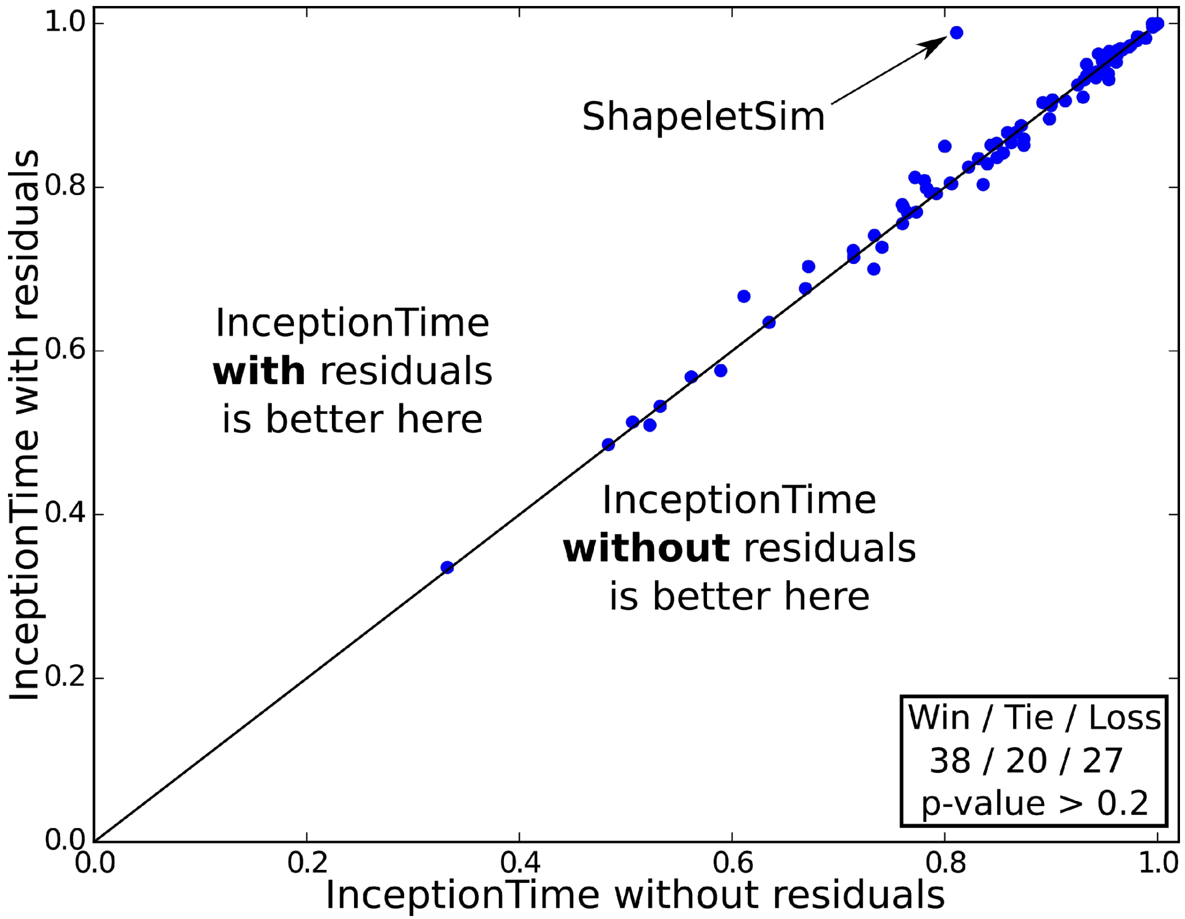
- InceptionTime significantly outperforms ResNet(5) [1]
- For DiatomSizeReduction the main improvement is from using a batch size larger than 1 (which is the case for the ResNet model for this specific dataset)



1. Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. (2019). Deep neural network ensembles for time series classification. IEEE International Joint Conference on Neural Networks.

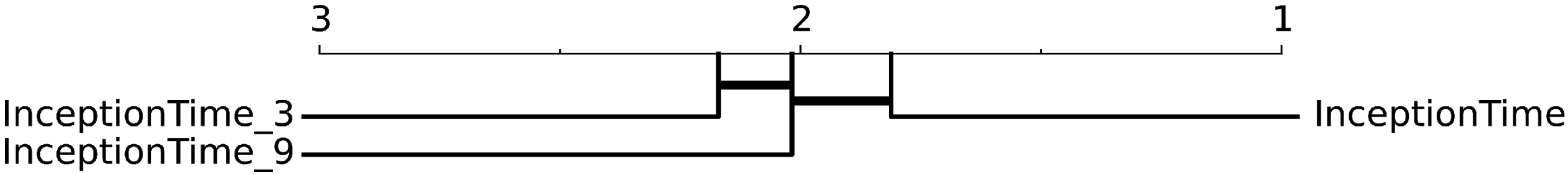
Hyperparameter study: Bottleneck & residual





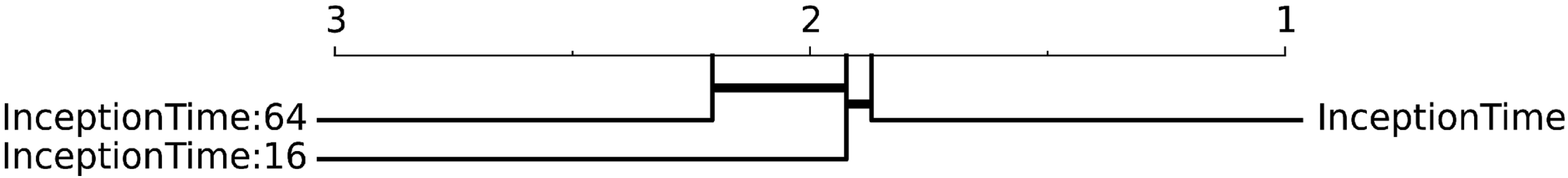
Further investigations ShapeletSim indicated that InceptionTime without the residual connections suffered from a severe overfitting.

Hyperparameter study: depth



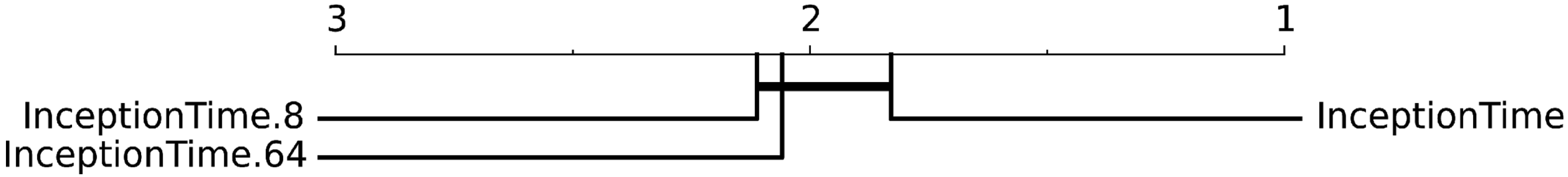
- InceptionTime_*x* denotes an InceptionTime with *x* layers
- InceptionTime is equivalent to InceptionTime_6 (the default value)
- A shallower model significantly decreases the accuracy
- A deeper model slightly decreases the accuracy
- Therefore we chose to use a network with 6 layers

Hyperparameter study: number of filters



- InceptionTime:*x* denotes an InceptionTime with *x* filters per module
- InceptionTime is equivalent to a model with 32 filters (default value)
- More filters showed a significant decrease in accuracy
- Less filters showed a slight decrease in accuracy
- This hyperparameter affects significantly the complexity of the model

Hyperparameter study: filter length



- InceptionTime.*x* denotes a model with a filter length equal to *x*
- InceptionTime is equivalent to a model with a filter length equal to 32
- The default value (32) showed a slight advantage
- Although larger values will produce a larger RF, these experiments showed that this hyperparameter should be carefully chosen

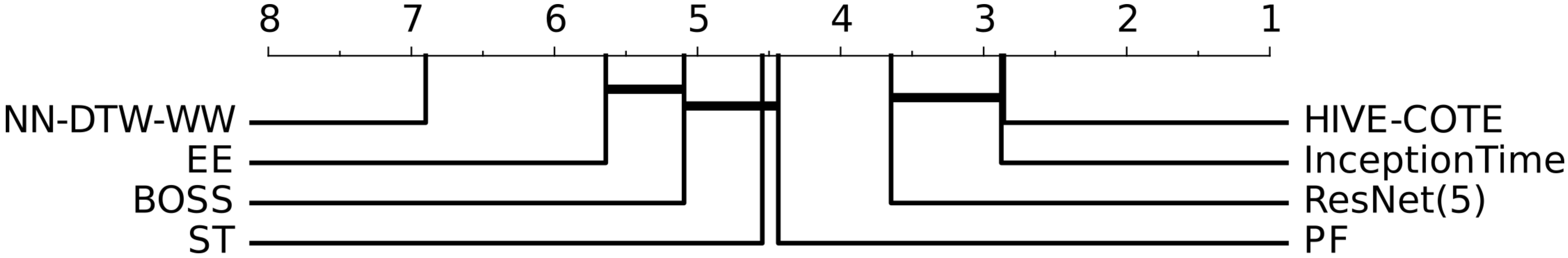
Receptive Field (RF) of a neural network

$$1 + \sum_{i=1}^{d} (k_i - 1)$$

- *d* represents the depth of the network
- k_i represents the length of the filters in i^{th} layer
- The stride is considered to be equal to 1
- RF can then be increased by either controlling *d* or k_i

For images, a large RF is needed to capture more context [1]

[1] Luo, W., Li, Y., Urtasun, R., & Zemel, R. (2016). Understanding the effective receptive field in deep convolutional neural networks. In *Advances in neural information processing systems*.


8
7
6
5
4
3
2
1
NN-DTW-WW
EE
BOSS
ST
HIVE-COTE
InceptionTime
ResNet(5)
PF