August 27-31, 2023
FLORENCE - ITALY
ICWE 16
16th International Conference on Wind Engineering

# Improving the Detection of Shapelets in Time Series for Thunderstorm Classification

M. Geuzaine [1], M. Arul [2], A. Kareem [3]

[1]*University of Notre Dame, Notre Dame, USA, mgeuzai4@nd.edu*
[2]*Virginia Tech, Blacksburg, USA, marul@vt.edu*
[3]*University of Notre Dame, Notre Dame, USA, kareem@nd.edu*

**SUMMARY:**
A tremendous amount of data related to wind has been recorded over the last decade in Mediterranean ports in order to understand the effects of thunderstorm winds on civil engineering structures. Automated classification techniques have thus been developed to detect these events of interest in such large databases. To ensure the autonomy and interpretability of the process, it is convenient to use a machine learning classifier trained on shapelet transforms. But the current algorithm is not able to identify a number of true positives and uses a lot of computational power. New techniques (Wavelet Decomposition, Randomized Sampling and Ensemble Classifier) are hence introduced in this paper to solve these problems.

*Keywords: shapelet transform, thunderstorm, classification, machine learning*

## 1. INTRODUCTION

To gain a better understanding of the devastating effects that thunderstorm winds can have on civil engineering structures, more particularly in Mediterranean ports, extensive measurement campaigns have been carried out during the last decade (Solari et al. 2012; Burlando et al. 2018). To do so, numerous monitoring systems and stations have been installed to record high-dimensional wind field measurements in a continuous way and with a high sampling rate. Given the tremendous amount of data that they generated over the years, it has become necessary to develop automated methods dedicated to detecting events of interest like thunderstorms from the analysis of time series.

Several techniques existed before (De Gaetano et al. 2014), but they often required the intervention of expert judgement through a detailed visual inspection of the time series to compensate for the absence of some statistics on which they are based. From a big data perspective, the use of machine learning has therefore appeared as a potential solution to ensure the autonomy of the process. Two main techniques have emerged since then. (Chen and Lombardo 2020) used a one-dimensional convolutional neural network classifier trained on segmented records while (Arul and Kareem 2021) used a random forest classifier trained on shapelet-transformed signals.

Shapelets are highly discriminative sequences that are discovered in the time series and that somehow represent the respective signature of any wind state. In a nutshell, these shapelets have first to be identified in the training dataset and can then be compared to the testing dataset to classify the recordings into two or more groups. By detecting local or global similarities in time series, this method reproduces what humans would naturally do when visualizing the measurements and thus has the advantage of being easily interpretable. In (Arul and Kareem 2022), this method also identified more thunderstorms than the above-mentioned statistical approaches and divided data into three categories, including intermediate events, instead of two.

However, despite the already high accuracy of the algorithm, some recordings were no longer considered true positives for thunderstorm events although they were before. In addition, and in general, the discovery of shapelets is also an extremely consuming step in terms of computational power when performed using brute force. To solve these problems, new techniques are implemented in this paper. Overall, the proposed method still relies on Shapelet Transforms (ST) but significant differences are introduced in the algorithm. They are discussed and illustrated in the next section. Meanwhile, the results obtained on the Wind and Ports datasets will be provided in an extended version of the present paper.

## 2. DIFFERENCES WITH THE PROPOSED APPROACH

In the first stage, the classification process starts by creating subsequences of the time series that are candidate shapelets. The similarity between these parts of signals is then measured and the quality of each shapelet is finally assessed. In the second stage, the minimum distance between the best shapelets and each time series is calculated and serves to train a machine learning classifier. This is the global methodology in which the following techniques are now introduced.

### 2.1. Wavelet Decomposition

In order to have a better understanding of the patterns associated with each time and frequency scale, a discrete wavelet transform is used to decompose the records into one low-frequency approximation component and a few high-frequency detailed components, as shown in Figure 1. This process accordingly divides the classification process into several sub-tasks, i.e. one for each component, and is supposed to improve its overall performance (Yan et al. 2020).

### 2.2. Randomized Sampling

Instead of considering all subsequences of all time series in the training dataset, a random number of them is skipped at each iteration (Renard et al. 2015). The probability distribution of this number is chosen to ensure that the pool of candidate shapelets remains diversified but is drastically reduced in size. The computational time needed to discover shapelets is hence expected to decrease by the same amount as well. This is schematically represented in Figure 2.

### 2.3. Ensemble Classifier

The two previous techniques are typically able to create several classifiers as they are or can respectively be trained on the transforms obtained from different sets of shapelets. It is necessary to perform an additional task which consists in reunifying their results, as represented in Figure 3. By combining the various views that the classifiers have on the data, the resulting algorithm is more stable and provides a more accurate decision.

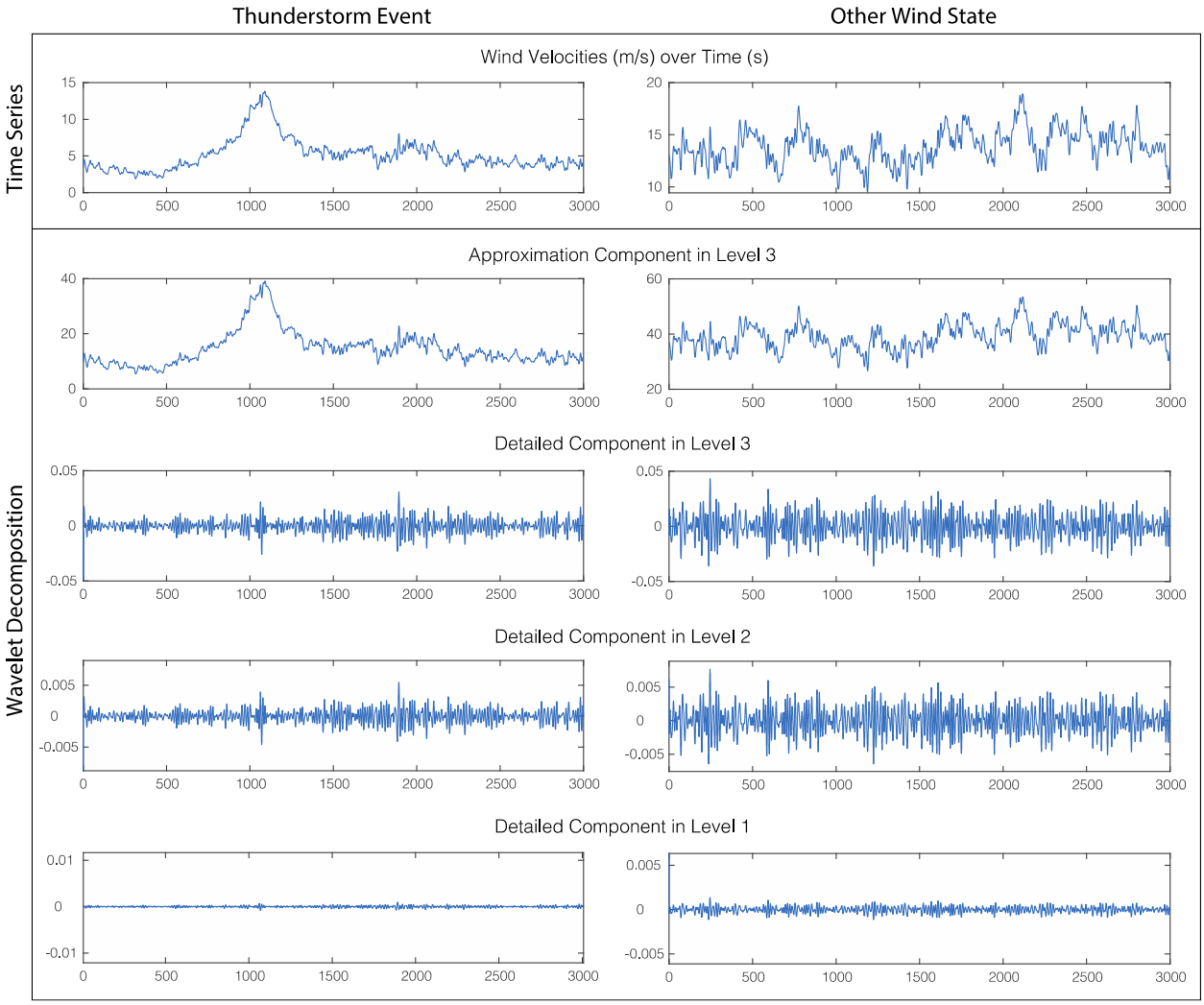


**Figure 1.** Wavelet Decomposition (db2) – Livorno, Anemometer 4, October 2 at 1PM and 15 at 2AM, 2015).

## 3. RESULTS AND CONCLUSIONS

This new algorithm will eventually be applied to classify the data recorded during the "Wind and Ports" project by ultrasonic anemometers in the Mediterranean ports of Genoa, La Spezia, Livorno and Savona in Italy, as well as Bastia in France. The results will be compared to the previous ones to evaluate the improvements due to the addition of new features, as detailed hereabove, in using less computational power and detecting more true positives.

Given that the proposed procedure requires training more classifiers, depending on the number of levels adopted for the wavelet decomposition, but creates fewer shapelets candidates due to the randomized sampling, depending on its probabilistic description, the reduction of the overall computational time obtained while varying these main parameters will be extensively discussed.

**ACKNOWLEDGEMENTS**

This paper is dedicated to the memory of Prof. Giovanni Solari and his dedication to the area of thunderstorm winds under the project THUNDERR. The first author acknowledges the financial support of a postdoctoral grant awarded by the Francqui Foundation, as part of the Belgian American Educational Foundation.

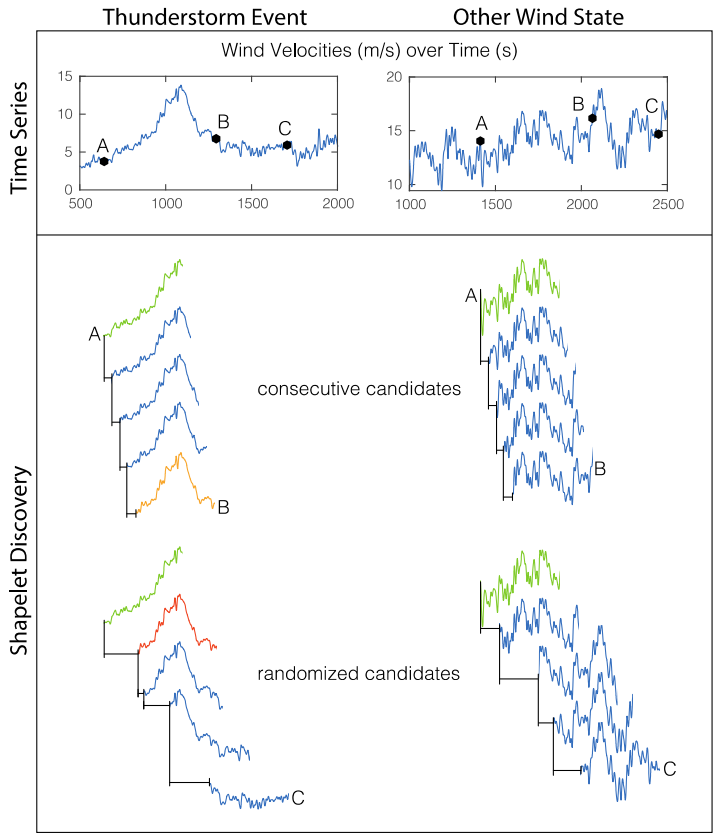


**Figure 2.** Randomized Sampling.

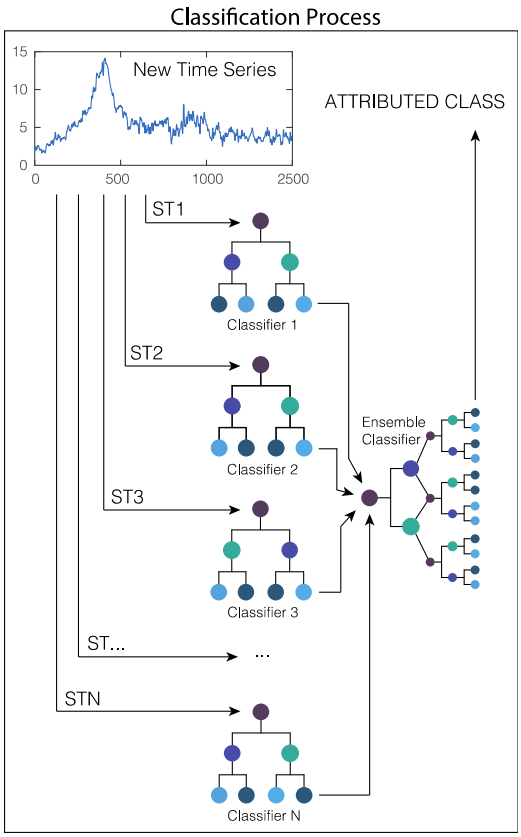


**Figure 3.** Ensemble Classifier.

## REFERENCES


Arul, Monica, and Ahsan Kareem. 2021. "Applications of Shapelet Transform to Time Series Classification of Earthquake, Wind and Wave Data." *Engineering Structures* 228: 111564.

Arul, Monica, Ahsan Kareem, Massimiliano Burlando, and Giovanni Solari. 2022. "Machine Learning Based Automated Identification of Thunderstorms from Anemometric Records Using Shapelet Transform." *Journal of Wind Engineering and Industrial Aerodynamics* 220: 104856.

Burlando, Massimiliano, Shi Zhang, and Giovanni Solari. 2018. "Monitoring, Cataloguing, and Weather Scenarios of Thunderstorm Outflows in the Northern Mediterranean." *Natural Hazards and Earth System Sciences* 18 (9): 2309–30.

Chen, Guangzhao, and Franklin T. Lombardo. 2020. "An Automated Classification Method of Thunderstorm and Non-Thunderstorm Wind Data Based on a Convolutional Neural Network." *Journal of Wind Engineering and Industrial Aerodynamics* 207: 104407.

Gaetano, Patrizia De, Maria Pia Repetto, Teresa Repetto, and Giovanni Solari. 2014. "Separation and Classification of Extreme Wind Events from Anemometric Records." *Journal of Wind Engineering and Industrial Aerodynamics* 126: 132–43.

Renard, Xavier, Maria Rifqi, Walid Erray, and Marcin Detyniecki. 2015. "Random-Shapelet: An Algorithm for Fast Shapelet Discovery." *Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015*, 1–10.

Solari, Giovanni, Maria Pia Repetto, Massimiliano Burlando, Patrizia De Gaetano, Marina Pizzo, Marco Tizzi, and Mattia Parodi. 2012. "The Wind Forecast for Safety Management of Port Areas." *Journal of Wind Engineering and Industrial Aerodynamics* 104–106: 266–77. https://doi.org/10.1016/j.jweia.2012.03.029.

Yan, Lijuan, Yanshen Liu, and Yi Liu. 2020. "Application of Discrete Wavelet Transform in Shapelet-Based Classification." *Mathematical Problems in Engineering* 2020.