




Time series extrinsic regression

Predicting numeric values from time series data

Chang Wei Tan¹  · Christoph Bergmeir¹ · François Petitjean¹ ·
Geoffrey I. Webb¹

Received: 22 June 2020 / Accepted: 17 February 2021 / Published online: 11 March 2021

© The Author(s), under exclusive licence to Springer Science+Business Media LLC, part of Springer Nature 2021

Abstract

This paper studies time series extrinsic regression (TSER): a regression task of which the aim is to learn the relationship between a time series and a continuous scalar variable; a task closely related to time series classification (TSC), which aims to learn the relationship between a time series and a categorical class label. This task generalizes time series forecasting, relaxing the requirement that the value predicted be a future value of the input series or primarily depend on more recent values. In this paper, we motivate and study this task, and benchmark existing solutions and adaptations of TSC algorithms on a novel archive of 19 TSER datasets which we have assembled. Our results show that the state-of-the-art TSC algorithm Rocket, when adapted for regression, achieves the highest overall accuracy compared to adaptations of other TSC algorithms and state-of-the-art machine learning (ML) algorithms such as XGBoost, Random Forest and Support Vector Regression. More importantly, we show that much research is needed in this field to improve the accuracy of ML models. We also find evidence that further research has excellent prospects of improving upon these straightforward baselines.

Keywords Time series · Regression · Machine learning

Responsible editor: Eamonn Keogh.

✉ Chang Wei Tan
chang.tan@monash.edu

Christoph Bergmeir
christoph.bergmeir@monash.edu

François Petitjean
francois.petitjean@monash.edu

Geoffrey I. Webb
geoff.webb@monash.edu

¹ Faculty of Information Technology, Monash University, 25 Exhibition Walk, Melbourne, VIC 3800, Australia

1 Introduction

In the past decade, there has been an increasing interest in time series analysis research, in particular time series classification (TSC) (Bagnall et al. 2017; Dau et al. 2019; Bagnall et al. 2015; Fawaz et al. 2019; Dempster et al. 2020; Tan et al. 2020b) and time series forecasting (TSF) (Hyndman 2018; Makridakis et al. 1982; Makridakis and Hibon 2000; Makridakis et al. 2018, 2020). TSC is the task of predicting a discrete label that classifies the time series into some finite discrete categories (Bagnall et al. 2017; Dau et al. 2019). On the other hand, TSF aims to predict future values of a series based on recent or seasonal values. It typically assumes that future values will more closely resemble recent values than those in the distant past.

Despite the thousands of papers published in both of these fields each year, there has been little investigation of *Time Series Extrinsic Regression* (TSER), i.e. a task to predict numeric values that depend on the whole series, rather than depending more on recent than past values such as TSF. The difference between TSC and TSER is that TSC maps a time series to a finite set of discrete labels while TSER predicts a continuous value from the time series. For instance, TSC might classify an ECG signal as arrhythmia or normal, while TSER could be used to predict a quantitative value such as the heart rate or respiratory rate of a patient (Pimentel et al. 2015, 2016; Meredith et al. 2012; Karlen et al. 2010) based on patterns in the ECG signal. TSER can be considered a special case of *scalar-on-function regression* (SoFR) from the statistics community (Reiss et al. 2017; Goldsmith and Scheipl 2014), where the functional data is a time series. SoFR considers a time series as functional data and builds statistical models to map functional data to a scalar response value. In our case, we address the problem from a ML perspective, treating it as a *regression* problem, taking time series data as the input and outputting a numeric value.

The term *regression* has different meaning in different contexts. In the broader machine learning context, *regression* means predicting a continuous numerical value from a set of features (Segal 2004; Sammut and Webb 2011). With respect to TSF, *regression* usually means fitting the historical time series data with a regression model such as ARIMA (Box and Jenkins 1970) or Exponential Smoothing (Gardner Jr 1985; Hyndman et al. 2008; Chatfield 1978) models to forecast future values of the time series. These TSF regression models typically heavily rely on recent or seasonal values, or sliding input windows of some form.

In this work, we refer to the TSER problem as a more general methodology of *predicting a single continuous scalar value from a time series*. We aim to predict values that can be either a continuation of the input time series or external to it and do not necessarily need to be a future value or depend on recent values. In the case where predicting a future value of a series is of interest, then that becomes a TSF problem. If predicting a finite discrete value is of interest, then that becomes a TSC problem. We are interested in a more general task that lies in between the spectrum of these two tasks, which cannot be solved intuitively using models from these two tasks or SoFR.

For instance, we are interested in predicting the heart rate of a person from accelerometer data (Reiss et al. 2019; Zhang et al. 2014), predicting the crop yield or fuel load from satellite image series describing the evolution of the ‘colours’ of the vegetation over the years; neither of which are discrete or future values. Figure

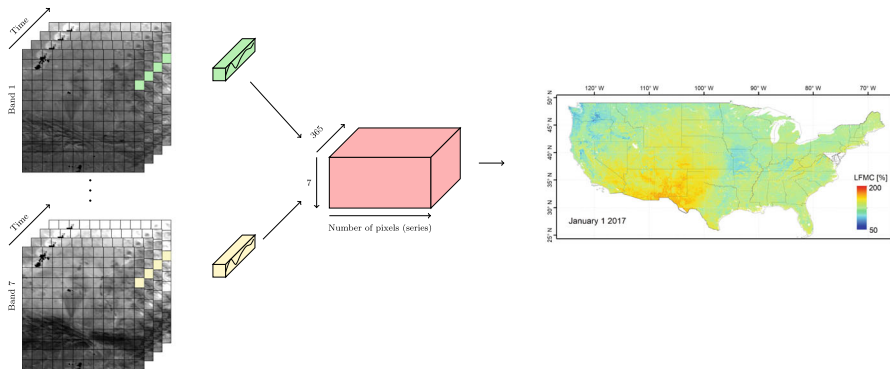


Fig. 1 Prediction of live fuel moisture content (LFMC) using satellite images time series

1 shows the example of predicting live fuel moisture content (LFMC) of the United States using a series of satellite images where the value of LFMC is a continuous value in the range from 0 to 200%. The input is the series of spectral values (i.e. time series of colour values) representing the state of a surface (or ‘pixel’) over the last 12 months; the target is to infer the amount of moisture in the vegetation, i.e. the ratio between the weight of water in vegetation and the weight of the dry part of vegetation (information that is obtained by sampling vegetation in the field, weighing it and drying it to weigh it again). This is a very important variable, as the risk of fire increases very rapidly as soon as the LFMC goes below 80% (Yebra et al. 2018), making it an invaluable variable for forest fire early warning systems. A very similar application is the one of predicting crop yield from these same series of spectral values, with great importance for food safety and agricultural planning (Pelletier et al. 2019).

Typical regression algorithms do not work well when applied directly to such problems because they do not take into account the temporal aspect of the data. These algorithms also suffer from the curse of dimensionality, especially when the data is sampled with high sampling frequency and with a large number of channels. TSC algorithms on the other hand were not designed for these continuous scalar outputs. In particular, they are predicated on the assumption that the output values are not ordered. Hence, we need algorithms that are able to learn the relationship between time series data and the continuous scalar variable. There has been some research in this area where the algorithms and features are specifically designed for the specific tasks (Reiss et al. 2019; Zhang et al. 2014; Zhang 2015; De Vito et al. 2008). Unfortunately, these algorithms do not generalise well to other problems. For instance, those specific features created from photoplethysmogram (PPG) measurements (Zhang et al. 2014; Reiss et al. 2019) for heart rate estimation cannot be used to predict crop yields and vice-versa.

Therefore in this paper, we aim to motivate the research into developing more general TSER algorithms. We start by introducing the first TSER benchmarking archive, which we have assembled and contains 19 datasets in various domains in Tan et al. (2020a). These datasets have varying number of dimensions, dimensions with unequal lengths and missing values. They are used to benchmark some adaptations of classical

regression and TSC algorithms as well as SoFR techniques. Our results show that simple variants of some state-of-the-art TSC algorithms outperform standard regression techniques (i.e. ones developed for tabular data) that do not take into account the underlying series nature of the data. More importantly, we show that most methods obtain similar accuracies and the top method—Rocket—is actually not far in accuracy from algorithms that ignore the sequential information in the series data, XGBoost (Chen and Guestrin 2016) and Random Forest (Breiman 2001), which motivates the need for the development of a subfield of research.

The rest of this paper is organised as follows. In Sect. 2, we introduce the problem that we aim to address and discuss the related work. Then we describe some of the applications of TSER with respect to the benchmark datasets we created in Sect. 2.2. Section 3 then describes how the classic regression and TSC algorithms can be adapted for TSER. After that, we evaluate these algorithms on the first TSER benchmark datasets in Sect. 4. Finally, in Sect. 5, we summarise our contribution and give some direction for future work.

2 Time series extrinsic regression

Time Series Extrinsic Regression (TSER) is a regression task that learns the mapping from time series data to a scalar value. It shares resemblance to other fields such as SoFR and time series regression, which has different meaning in different contexts. In this section, we give a formal definition to TSER as we employ it. We will also try to clear any misunderstandings that the readers might have and introduce the task that we aim to address. We first define a time series in Definition 1.

Definition 1 A time series S is an ordered collection of L pairs of measurements and timestamps, $S = \{(s_1, t_1), (s_2, t_2), \dots, (s_L, t_L)\}$, where $s_i \in \mathbb{R}^D$ and t_1 to t_L are the timestamps for some measurements s_1 to s_L .

Note that the D -dimensional measurement s_i measures the same phenomena with different instruments at the same time. Time series data differs from static data in a way that the ordering of the data attribute in time series data is critical in finding the best discriminating features in time series data.

Classification and *Regression* are both supervised learning tasks that learn the relationship between a target variable and a set of features (Sammut and Webb 2011). The main difference between *Classification* and *Regression* is that *Classification* predicts a categorical value for a data instance that categorises the data into some finite categories, while *Regression* predicts a continuous value. *Regression* tasks can become *Classification* tasks when the predicted values are discretized into some finite labels for the data. In this work, we only focus on *Regression*. For example, the simplest regression algorithm, linear regression, assumes a linear relationship between a set of predictors (features) and a target variable, and fits a straight line through all the predictors to generate a prediction for the target variable.

Traditionally in ML, the features used for regression are static and have no relation to time. For instance, we could predict house prices using features such as the number of bedrooms, crime rate, nitric oxides concentration (pollution level), accessibility

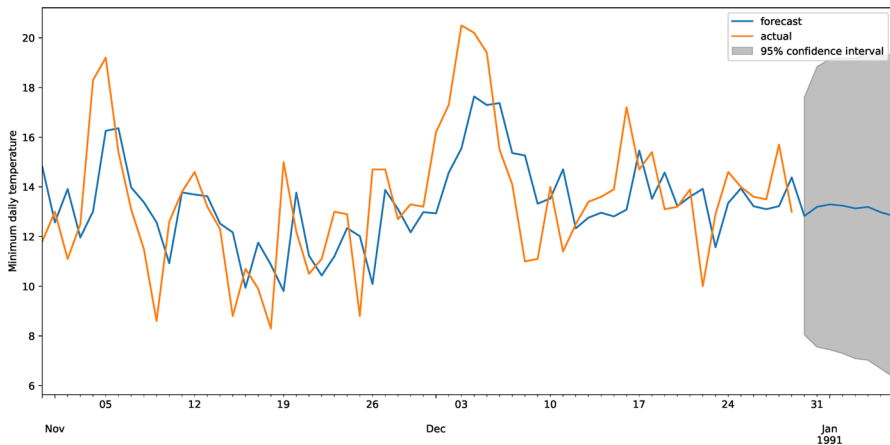


Fig. 2 Example of an autoregression model of order 7, AR(7)

to radial highways and weighted distances to employment centers.¹ These features (predictors) do not depend on time and are less likely to change over time. They are then used to train an ML algorithm such as a Random Forest (Breiman 2001), XGBoost (Chen and Guestrin 2016) or even linear regression to predict house price, the target variable that we are interested in. Different from the traditional regression problem, the regression problem that we tackle in this work, considers time series data as the features. With respect to the house price prediction example, instead of using a single value for the number of rooms, crime rate or pollution level, we use the time series of these features to predict house prices. For example the daily crime rate or daily pollution level over the last one month. A more concrete example of regression in our context is the prediction of heart rate which can only be achieved using time series data such as PPG and accelerometer data (Reiss et al. 2019; Zhang 2015; Zhang et al. 2014) that measures the pulse and movement of the subject within a certain period of time.

A very large branch of time series analysis deals with TSF (Hyndman 2018; Hyndman et al. 2008; Makridakis et al. 2018), where *regression* carries a slightly different meaning. In TSF, *regression* is used to fit autoregressive models on the historical time series which models the recent and/or seasonal values in the time series. Figure 2 shows an example of a linear autoregressive model of order 7, AR(7), i.e. the model uses the past 7 days minimum daily temperature to forecast the minimum daily temperature for the next day.

These models are then extrapolated to predict future values of the same time series. Going back to the example of predicting house prices, autoregressive models can be used to fit past house prices data and produce a good forecast for future house prices, as it is very likely that house price depends on the price in the previous months. In our regression context, we can also build models to predict future house price using past house prices. However, we aim at developing more general models that do not

¹ <https://www.kaggle.com/vikrishnan/boston-house-prices>.

make the assumptions that frequently underlie forecasting models, such as that the most recent values are most indicative of future values. In other words, we can see that forecasting models will not be useful in our regression example of predicting heart rate, as heart rate is not a future value of ECG, PPG and accelerometer signal and does not depend more on the final value of these data than on the initial ones.

Rather, heart rate is a quantitative value of the signal that can be obtained through counting the number of peaks in the signal. Formally, we define the task of *Time Series Extrinsic Regression* in Definition 2.

Definition 2 A *time series extrinsic regression model* is a function $\mathcal{T} \rightarrow \mathcal{R}$, where \mathcal{T} is a class of time series. *Time series extrinsic regression* seeks to learn a regression model from a dataset $\mathcal{D} = \{(t_1, r_1), \dots, (t_n, r_n)\}$, where t_i is a time series and r_i is a continuous scalar value.

2.1 Related work

Time series data can be considered as functional data, where the measurements are a function of time (Goldsmith and Scheipl 2014). Functional regression is a widely studied task in the statistics community (Reiss et al. 2017; Goldsmith and Scheipl 2014). Functional regression models can be classified into three categories: (1) scalar responses with functional predictors (scalar-on-function regression); (2) functional responses with scalar predictors (function-on-scalar regression); and (3) functional responses with functional predictors (function-on-function regression) (Reiss et al. 2017). The task of mapping a time series to a scalar value, TSER, is closely related to scalar-on-function regression (SoFR), a task that maps functional data (e.g., a time series) to a scalar response (Reiss et al. 2017; Goldsmith and Scheipl 2014). SoFR typically works by first representing the time series data in its functional form. Then a basis function such as Functional Principal Components (FPC), B-spline, Fourier or Wavelet can be applied to smooth the data and reduce noise. Finally a regression model is applied to the smoothed data to predict the scalar value.

Functional linear models (FLM) are the most common approach for SoFR, which extend the standard multiple linear regression model to functional data (Goldsmith and Scheipl 2014). Most work in the literature of SoFR focused on better estimating the weights that are applied to every timestep of the time series data (Goldsmith and Scheipl 2014). The study of Goldsmith and Scheipl (2014) shows that SoFR models have been applied to problems such as predicting annual rainfall from observed temperature and predicting fat content in meat from near-infrared spectrum. The study compares various SoFR models with its ensemble counterparts and non-functional models such as random forest and gradient boosting machines. The results concluded that ensembles of models work better than a single model. More importantly, the results also show the limitation of FLMs where non-functional models such as random forest are robust and consistently outperform other FLMs on all the test datasets. In addition, functional regression models usually require an in-depth understanding of the data on hand and experience, in order to apply the right basis function to fit the model. For instance, Fourier basis functions will not work well on non-periodic signals.

While we have not been able to identify any prior work in the ML community specifically addressing the more general class of learning task that we call *time series extrinsic regression*, there are a number of specialised techniques addressing specific cases. In addition to forecasting, one that has received considerable attention is heart rate (HR) estimation using photoplethysmogram (PPG) sensors (Reiss et al. 2019; Zhang et al. 2014). These methods rely on spectral analysis (Zhang et al. 2014; Zhang 2015; Salehizadeh et al. 2016; Schäck et al. 2017) but they were not very accurate (Reiss et al. 2019). A convolutional neural network based approach that takes the signal in the frequency domain as input has been proposed to improve the prediction accuracy (Reiss et al. 2019). This approach was shown to be significantly more accurate compared to the existing spectral methods.

Similar to heart rate estimation, respiratory rate (RR) estimation can also be achieved using PPG sensors (Pimentel et al. 2016; Meredith et al. 2012; Pimentel et al. 2015). Estimating RR is an important task because it is often the earliest sign of critical illness (Meredith et al. 2012). Existing methods fail to distinguish between periods of high and low quality data and were not able to generalise well to other datasets (Pimentel et al. 2016). Typically, estimation of RR from PPG is achieved by applying a moving window to the time series producing an estimate for RR per window (Pimentel et al. 2016) and consists of four key components, (a) extracting respiratory signals; (b) estimating respiratory rates; (c) fusing the estimates and (d) quality assessments (Pimentel et al. 2015, 2016). A probabilistic approach was proposed (Pimentel et al. 2015) using the Gaussian process regression framework to extract RR from the different sources of modulation in the PPG signal. The authors then proposed another method (Pimentel et al. 2016) by fitting multiple autoregressive models to the extracted respiratory signals. Their method was evaluated on two datasets, the Capnobase (Karlen et al. 2010) and the BIDMC dataset (Pimentel et al. 2016) (both can be found in <http://peterhcharlton.github.io/RRest/datasets.html>). Although the results showed that their method achieved the best mean absolute error (MAE) on both datasets compared to other existing methods in RR estimation, it was only significantly different to one of the methods on the Capnobase dataset. There were no significant difference on the BIDMC dataset.

Other than health monitoring, there are also similar works done for pollution monitoring, where the goal is to predict pollutant concentration using on-field sensors (De Vito et al. 2008). De Vito et al. (2008) proposed a simple feed-forward network with 5 hidden layers, taking 7 sensor inputs to estimate benzene concentration in an Italian city. The method, although simple, achieved very low MAE of $0.13\mu\text{g}/\text{m}^3$, but is not generalisable.

2.2 TSER applications and datasets

To support research TSER, we created the first TSER benchmarking archive, available online at <http://tseregression.org/>. In this section, we describe the possible applications of TSER and our first TSER archive. The current TSER archive contains 19 time series datasets from 5 application areas, *Health Monitoring*, *Energy Monitoring*, *Environment Monitoring*, *Sentiment Analysis* and *Forecasting*. The archive contains 8 datasets

assembled from the UCI machine learning repository (Dua and Graff 2017), 3 from physionet.org, 1 from a signal processing competition (Zhang et al. 2014), 1 from the Covid-19 database from the World Health Organisation, 1 from the Australian Bureau of Meteorology (BOM) and the rest are donations. These datasets are unnormalised with varying number of dimensions, unequal length dimensions and missing values. We briefly describe these datasets below and refer readers to Tan et al. (2020a) for a more detailed description. Table 1 outlines the properties of the datasets in the current TSER archive.

2.2.1 Energy monitoring

With advances in Smart City and Internet of Things applications, the task to monitor energy and power consumption has become more important than ever. The ability to predict energy and power consumption accurately can save millions of dollars for a big company. Energy monitoring is typically done by collecting data such as temperature, humidity, rain, voltage and current readings from sensors attached all over a building. These data are collected in the form of time series and is mapped to the power consumption of the building. For example, higher power consumption will be observed during winter months as more energy is required to heat up a building. The **AppliancesEnergy**, **HouseholdPowerConsumption1** and **HouseholdPowerConsumption2** are the three datasets in this archive targeting this application. Figure 3 shows an example of time series data in the HouseholdPowerConsumption datasets.

2.2.2 Environment monitoring

In the context of climate change, environment monitoring has become more important than ever. Environment monitoring is the task of predicting anything related to our environment such as pollution level, rainfall, crop yield and flood water level. The three datasets **BenzeneConcentration**, **BeijingPM10Quality** and **BeijingPM25Quality** focus on predicting pollution level in a metropolitan city. The **LiveFuelMoistureContent** is a dataset about predicting live fuel moisture content (moisture content in leaves) using series of satellite images, which we described in the introduction. Predicting the moisture content is very critical in bushfire prevention that could prevent the lost of thousands of lives and millions to billions of dollars. Figure 4 shows an example of the satellite image time series of a tree cover with 7 spectral bands in the LiveFuelMoistureContent dataset. The three **FloodModeling** datasets address prediction of the height of different riverbeds given a series of rainfall events. Here again, being able to predict the rise of water is critical to mitigate its risk. The relationship between rainfall and water height in different locations is non-linear, as it depends on topography, transpiration and rainfall dynamics. Here we assume that topography and land-cover (which drives transpiration) is not known and propose to model water height directly from rainfall time series. Finally, the **AustraliaRainfall** dataset contains the hourly temperature of various locations in Australia and the goal is to predict the total daily rainfall in those locations based on the hourly temperature. This is useful as temperature sensors are much cheaper and easy to maintain as compared to rain gauges.

Table 1 Time series datasets in the current TSER archive

Id	Dataset	Train size	Test size	Length	No. of dimension	Missing
<i>Energy monitoring</i>						
1	AppliancesEnergy	96	42	144	24	No
2	HouseholdPowerConsumption1	746	694	1440	5	Yes
3	HouseholdPowerConsumption2	746	694	1440	5	Yes
<i>Environment monitoring</i>						
4	BenzeneConcentration	3433	5445	240	9	Yes
5	BeijingPM25Quality	12,432	5100	24	9	Yes
6	BeijingPM10Quality	12,432	5100	24	9	Yes
7	LiveFuelMoistureContent	3493	1510	365	7	No
8	FloodModeling1	471	202	266	1	No
9	FloodModeling2	389	167	266	1	No
10	FloodModeling3	429	184	266	1	No
11	AustraliaRainfall	112,186	48,081	24	3	No
<i>Health monitoring</i>						
12	PPGDalia*	43,215	21,482	256 & 512	4	No
13	IEEPPG	1768	1328	1000	5	No
14	BIDMCRR	5471	2399	4000	2	No
15	BIDMCHR	5550	2399	4000	2	No
16	BIDMCSpO2	5550	2399	4000	2	No
<i>Sentiment analysis</i>						
17	NewsHeadlineSentiment	58,213	24,951	144	3	No
18	NewsTitleSentiment	58,213	24,951	144	3	No
<i>Forecasting</i>						
19	Covid3Month	140	61	84	1	No

The ones marked with an asterisk (*) have different lengths from one dimension to another (but the length is the same for all instances in any single dimension)

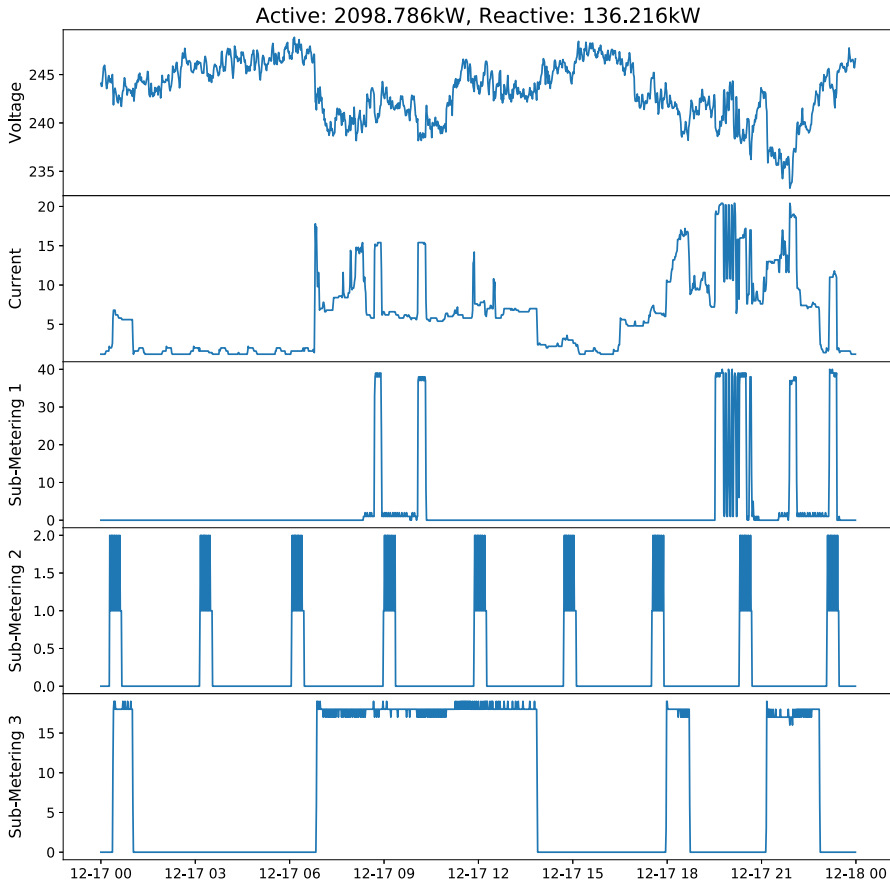


Fig. 3 Examples of the daily voltage, current and sub-metering measurements in the HouseholdPowerConsumption dataset that is used to predict the total daily active and reactive power consumption in a house

2.2.3 Health monitoring

Health monitoring is the task of monitoring the health or vital signs of an individual. The data typically comes from a wearable device that can be attached to the subject, such as a photoplethysmogram (PPG), electrocardiogram (ECG), electroencephalogram (EEG) or accelerometer. In this work, we focus on three tasks, estimating heart rate, respiratory rate and blood oxygen saturation level. The **PPGDalia**, **IEEPPG** and **BIDMCHR** are datasets focusing on heart rate estimation. Figure 5 illustrates an example of the PPG and accelerometer signal from the PPGDalia dataset. **BIDMCRR** and **BIDMCSpO2** are both datasets on predicting respiratory rate and blood oxygen saturation level, respectively.

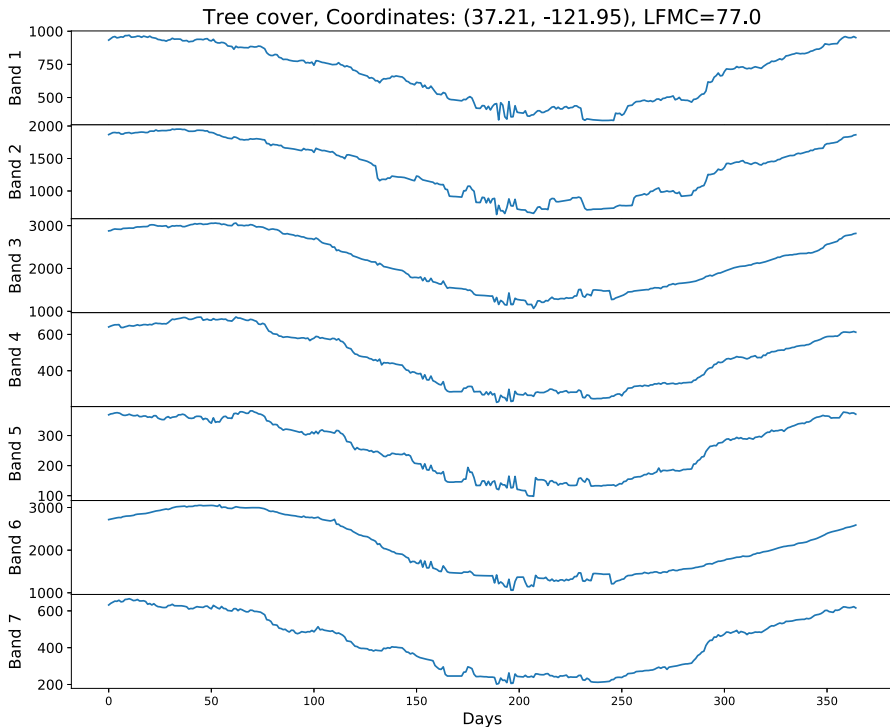


Fig. 4 Example of LiveFuelMoistureContent time series with 7 spectral bands

2.2.4 Sentiment analysis

Sentiment analysis is the interpretation and classification of emotions (positive, negative or neutral) within some text using text analysis techniques. This is typically done by analysing text comments or posts on websites and social media platforms to predict a sentiment score (Moniz and Torgo 2018). Moniz and Torgo (2018) released a dataset containing 100,000 news items on four topics: *economy*, *microsoft*, *obama* and *palestine* with the respective social feedback on 3 social media platforms: *Facebook*, *Google+* and *LinkedIn*. Here we attempted a different approach to predict the sentiment score by analysing the number of reactions received for the piece of news on the respective social media platforms. We included the **NewsHeadlineSentiment** and **NewsTitleSentiment** datasets that aim to predict the sentiment score of news headline and news title using the number of reactions over time from social media platforms illustrated in Fig. 6.

2.2.5 Forecasting

As described in the introduction and Sect. 2, TSF is the task of predicting future values based on some recent and/or seasonal values. This is usually done by fitting a model to the historical data and extrapolating it into the future. Our regression problem can be

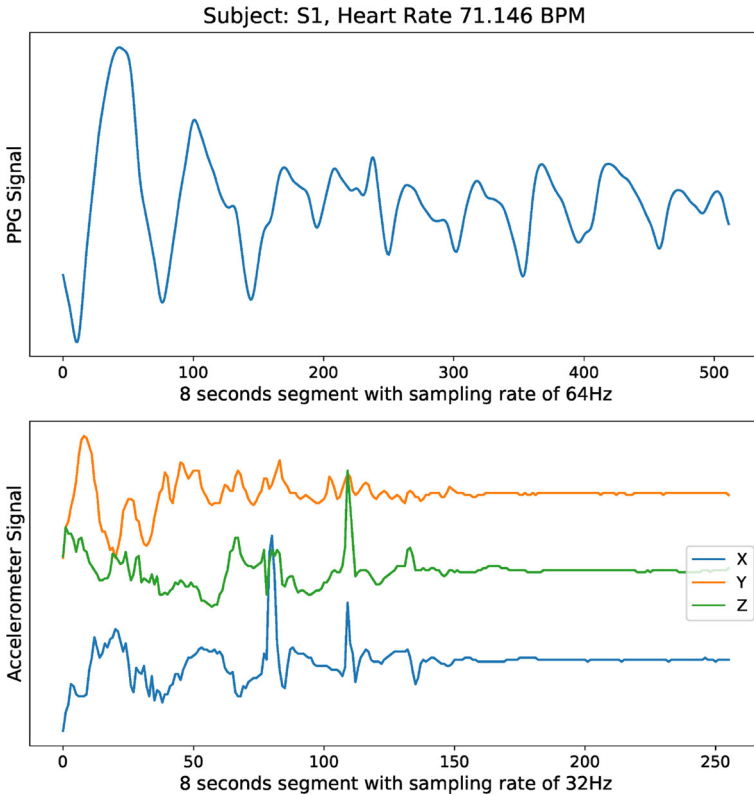


Fig. 5 Example of time series in the PPGDalia dataset. The title shows the subject and the current heart rate in beats per minute (BPM)

seen as a general case of forecasting where we are still predicting a continuous value that may not necessarily be a future value or depending more heavily on recent values. Thus, we included in this archive a dataset that could easily be solved with forecasting models to show that forecasting tasks can also be tackled using TSER models. The **Covid3Month** dataset contains the daily confirmed number of COVID-19 cases in most countries in the world from January to March 2020, and the goal is to predict the death rate at the start of April 2020. An example of the daily confirmed Covid-19 cases and death rate for Italy is shown in Fig. 7.

3 Existing algorithms

In this section, we describe some existing algorithms for TSER problems. Most methods developed in TSER cases are highly specific to a problem and are not generalisable, as discussed in Sect. 2.1. We observe the similarity of TSER with TSC (Bagnall et al. 2017) in Definition 2. The only difference between both tasks is that the target variable is continuous instead of discrete for TSC. Hence, in principle, most methods devel-

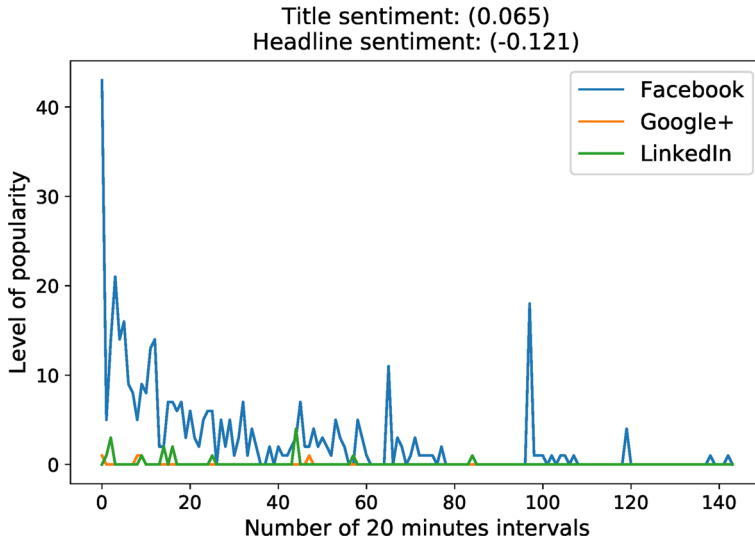


Fig. 6 Example of news popularity on 3 social media platforms. The title of the news is “Obama denounces rise of ‘vulgar and divisive’ politics of Trump” with the headline “And it’s worth asking ourselves what each of us may have done to contribute to this vicious atmosphere in our politics,” Obama told the ...” (<https://time.com/4259468/obama-trump-violence-rallies/>). The values in the brackets correspond to the respective sentiment value in news title and headline after 2 days

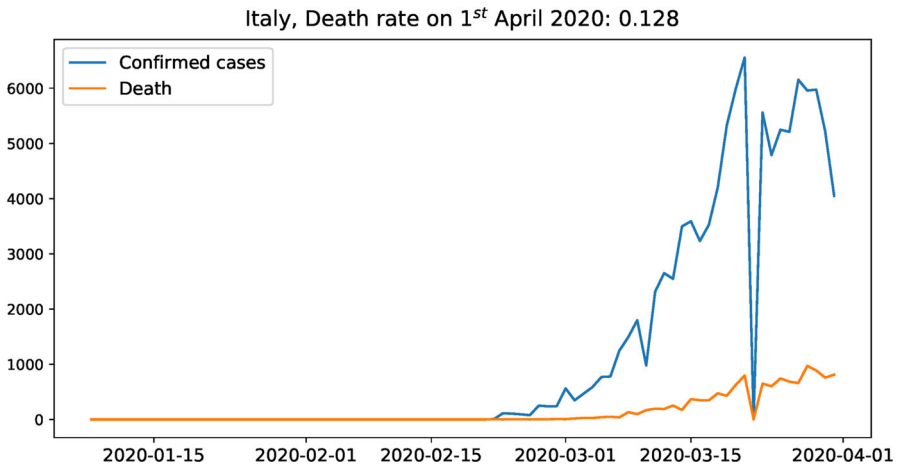


Fig. 7 Daily confirmed Covid-19 cases and death rate for Italy

oped for TSC can be adapted for TSER problems. These algorithms are categorised into 4 types: feature-based, dictionary-based, distance-based and deep learning.

3.1 Feature-based algorithms

Feature-based algorithms learn from time series data by extracting discriminating features. Then these features are used to train a classification or regression model. In this section, we discuss some existing feature-based algorithms for time series data.

3.1.1 Classical regression models

Classical regression models such as Support Vector Machine (SVM), Linear Regression (LR) and Random Forest (RF) are designed for tabular data. These models learn a mapping function from some input features extracted from the time series to the target variable. A straightforward approach is to treat the time series as tabular data where the time series values are the features. Multidimensional time series will be flattened out into a single long feature vector of length $D \times L$, where D is the number of dimensions in the series and L is the length of the time series. For instance, a time series with 3 dimensions and 100 data points results in a feature vector with 300 features. Generally, this approach will not take into account the temporal dimension which is important for discriminating time series because each feature is assumed to be independent of one another.

Despite the simplicity of treating the values of each time series as the features, a more common practice is to extract features from the whole time series. These features are used to characterise the time series which are commonly used for forecasting and visualisation (Kang et al. 2017; Montero-Manso et al. 2020). The FFORMA algorithm (Montero-Manso et al. 2020) is a feature-based forecast model that trains a meta-model using features extracted from the time series. The meta-model is used for assigning weights to various forecasting algorithms based on the characteristics of the time series. Features are also being used to visualise the performance of forecasting algorithms in an instance space, where time series are represented in a 2-dimensional space (Kang et al. 2017). These features include the summary statistics of the time series, spectral entropy, trend, seasonality, linearity and autocorrelation are extracted from the time series (Kang et al. 2017; Montero-Manso et al. 2020). The `tsfeatures` R package² is a popular package that extracts various features from time series data. Fulcher et al. (2013) introduce the Highly Comparative Time Series Analysis (HCTSA) features set that consists of over 7000 time series features. The Canonical Time Series Characteristics (Catch22) (Lubba et al. 2019) is a reduced set of HCTSA that consists of the 22 most discriminating features for TSC, evaluated on the UCR TSC archive. Although Catch22 when trained with a decision tree classifier is not as accurate as some state-of-the-art TSC algorithms, it is more interpretable, which may important in some applications.

Once the features are extracted, they can be used with any classical regression model. Next, we discuss some of the popular regression models. The SVM (Cortes and Vapnik 1995) is a popular classification model. Support Vector Regression (SVR, Drucker et al. 1997) is a variant of SVM designed for regression. Unlike many regression algorithms that seek to minimize squared error, SVR tries to fit the error rate

² <https://github.com/robjhyndman/tsfeatures>.

within a threshold, ϵ (Drucker et al. 1997). SVR works by mapping the data into a higher-dimensional space so that it is linearly separable using a kernel function such as linear, sigmoid or Gaussian Radial Basis Function (RBF, Cortes and Vapnik 1995). Then it fits a hyperplane through the data bounded by two boundary lines which are ϵ distance apart from the hyperplane. The boundary lines are formed by support vectors which are datapoints that are closest to the boundary.

The RF (Breiman 2001) algorithm has proven to be very robust on many classification and regression tasks (Segal 2004). It is a bootstrap aggregation (also known as bagging) ensemble learning method that combines the predictions of multiple decision trees to improve prediction accuracy (Breiman 2001). Bagging is a type of ensemble learning method that randomly samples the data with replacement to build multiple models and aggregates the outputs from all models. Bagging aims to reduce the variance of high variance models such as decision trees. RF builds a multitude of decision trees at training time and outputs the average values of the appropriate leaf for regression tasks (Breiman 2001). There are 2 main hyper-parameters that need to be tuned for each problem, the number of trees N_{tree} and the number of features randomly selected at each node m (Breiman 2001). One major disadvantage of RF is that it is prone to overfit datasets with noisy classification/regression tasks.

Extreme Gradient Boosting (XGBoost, Chen and Guestrin 2016) is a further accurate and popular machine learning algorithm. Similar to RF, XGBoost is a decision tree based ensemble learning algorithm that aims to reduce the variance and bias. Different from RF that uses bagging, XGBoost uses gradient boosting with regularisation to avoid overfitting, a problem in RF (Chen and Guestrin 2016). XGBoost reduces bias by building models sequentially while minimising the errors from previous models (Chen and Guestrin 2016). The errors are minimised using the gradient descent algorithm. This essentially “boosts” the model’s performance over time (Chen and Guestrin 2016).

3.1.2 Functional linear models

SoFR is widely studied in the statistics community. Specifically FLM is the most common approach for SoFR as it is simple and intuitive (Goldsmith and Scheipl 2014). FLM extends the standard multiple linear regression model to functional data (Goldsmith and Scheipl 2014). It is expressed as $Y_i = \int_0^1 W_i(s)\beta(s)ds + \epsilon_i$, where Y_i is the scalar response, $W_i(s)$ is the functional form of the time series, $\beta(s)$ is the coefficient function and ϵ_i is the random noise in the data (Reiss et al. 2017; Goldsmith and Scheipl 2014). Most work in the literature of SoFR had focused on better estimating the $\beta(s)$ coefficient function with various basis functions.

In this work, we will only be focusing on the two most popular basis functions for FLM. The FPC basis function when applied to FLM is commonly known as functional principal component regression (FPCR). FPCR is based on functional principal component analysis (FPCA) decomposition (Goldsmith and Scheipl 2014) that is similar to PCA decomposition where the data is represented by K_w principal components that explain the most variance in the data. Other than FPC, the smoothness in the coefficient function can be enforced using spline basis functions (Goldsmith and Scheipl

2014). The B-spline basis function is one of the most popular choices where the $\beta(s)$ coefficient function is expressed in terms of K_B B-spline basis.

3.1.3 Interval-based features

Instead of extracting features from the whole time series, features can be extracted from the intervals of the time series. It has been shown that these interval-based features generally give better performance than whole series features (Deng et al. 2013; Bagnall et al. 2017). The Time Series Forest algorithm (Deng et al. 2013) is one of the most accurate TSC algorithms (Bagnall et al. 2017). It extracts three features, mean, standard deviation and slope from an interval of a time series and builds a forest of time series trees, where random intervals are selected in each node of the tree (Deng et al. 2013).

Time series shapelets algorithms (Ye and Keogh 2009; Rakthanmanon and Keogh 2013; Lines et al. 2012) find the best discriminating shapelets (subsequences) in the data. The first time series shapelets classifier (Ye and Keogh 2009) trains a decision tree using shapelets as the splitting criterion. However, the algorithm has very high training complexity as it needs to scan through a high number of shapelet candidates. Since then, many novel scalable algorithms for shapelet discovery have been proposed (Rakthanmanon and Keogh 2013; Mueen et al. 2011; Grabocka et al. 2014; Lines et al. 2012). The most accurate shapelet algorithm, Shapelet Transform (ST) (Lines et al. 2012) transforms a time series using the distance of a time series to all k shapelets, creating a feature vector with k dimensions. The transformed time series are then used to construct one of the most accurate TSC algorithms, Shapelet Ensemble (SE) (Bagnall et al. 2015). SE is an ensemble consisting of 8 standard classifiers each applied to the shapelet features.

3.1.4 Random convolutional kernel transform (Rocket)

Recently, Dempster et al. (2020) proposed the Rocket classifier that achieves state-of-the-art accuracy in TSC with a fraction of the computational expense of existing methods. Rocket transforms time series using a large number of random convolutional kernels and trains a ridge regression classifier. These kernels have random length, weights, bias, dilation, and padding, and when applied to a time series produce a feature map. Then the maximum value and the proportion of positive values are computed from each feature map, producing two real-valued numbers as features per kernel. With the default 10,000 kernels, Rocket produces 20,000 features. Rocket was found to be the most accurate TSC classifier compared with other state-of-the-art algorithms such as HIVE-COTE (Lines et al. 2016) and InceptionTime (Fawaz et al. 2020) when benchmarked on the 85 TSC datasets (Dau et al. 2019). As Rocket was designed for classification tasks, in this work, we adapt Rocket for regression tasks by replacing the ridge regression classifier with a ridge regression model.

3.2 Dictionary-based algorithms

Dictionary-based algorithms transform time series by building a dictionary that represents the observed frequency of a particular pattern or feature in the time series. The algorithms then learn to discriminate between different time series by comparing the dictionary of the two time series. This is also known as the “bag of words” algorithm where the patterns (subsequences) are discretized into words.

There are various bag of words algorithms for TSC. Notably some of the popular ones are the Bag of Patterns (BOP) (Lin et al. 2012), Symbolic Aggregation Approximation Vector Space (SAXVSM) (Senin and Malinchik 2013), Bag of Symbolic Fourier Approximation (SFA) Words (BOSS) (Schäfer 2015), Word Extraction for TSC (WEASEL) (Schäfer and Leser 2017a) and WEASEL + Multivariate Unsupervised Symbols and Derivatives (MUSE) (Schäfer and Leser 2017b).

The recent TSC benchmark survey (Bagnall et al. 2017) ranks BOSS as the most accurate dictionary-based classifier. BOSS builds a dictionary using SFA words (Schäfer 2015). Each subsequence in the time series is transformed into SFA words using truncated discrete fourier transform, making it robust to noise.

Although the survey (Bagnall et al. 2017) did not compare with WEASEL, WEASEL is arguably more accurate than BOSS (Schäfer and Leser 2017a). WEASEL improves on BOSS by determining discriminative Fourier coefficients using ANOVA *f*-test and applying Chi-Squared test for feature selection (Schäfer and Leser 2017a). WEASEL+MUSE aims to tackle multivariate TSC by splitting the time series into its dimensions and applying the univariate transformation to each dimension (Schäfer and Leser 2017b). It also transforms the derivative of each dimension into words and concatenates these with a dimension identifier to enrich the feature space. Finally, similar to WEASEL, a feature selection technique is applied to filter out non-discriminative features (Schäfer and Leser 2017b).

3.3 Distance-based algorithms

The majority of TSC research has been focused on the similarity of two time series. This involves matching two time series and computing the distance between them. Then, a *k*-nearest neighbour (*k*-NN) algorithm is applied to find the most similar time series (Lines and Bagnall 2015; Tan et al. 2020b).

The *k*-Nearest Neighbour (*k*-NN) algorithm is one of the simplest and most intuitive algorithms (Sammut and Webb 2011). A *k*-NN algorithm requires two parameters, (1) the number of nearest neighbours *k* and (2) a distance metric (Sammut and Webb 2011). Similar to any other classical regression models described in Sect. 3.1.1, *k*-NN was initially designed for tabular data. Some examples of distance metrics for tabular data are the Euclidean, Manhattan, Minkowski and Mahalanobis distances. Using one of these distance metrics, the model finds *k* nearest instances from the training dataset to a query instance in the feature space (Sammut and Webb 2011). For regression, the target values of the *k* nearest neighbours are averaged out and assigned as the target of the query instance. Weighted average can also be applied using the distances to the query to put more emphasis on nearer neighbours.

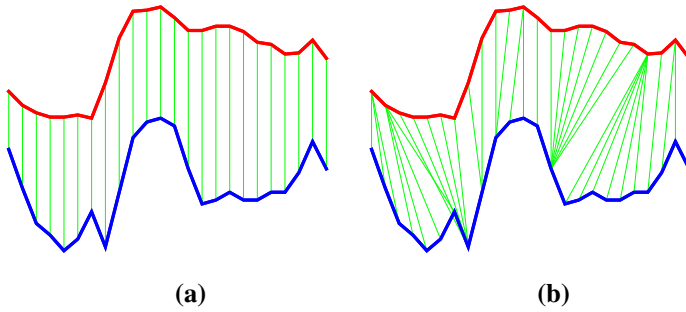


Fig. 8 Example of alignment of two time series using **a** Euclidean distance and **b** DTW distance

For time series data, the k -NN algorithm has to take into account the temporal dimension of the data. Hence, the distance measures (Lines and Bagnall 2015; Tan et al. 2020b) are also different from classic k -NN algorithms for tabular data. They are commonly known as elastic distances. The simplest is the Euclidean distance (ED), which is similar to the ED used for tabular data. Equation 1 describes the ED to compute the distance between two time series P and Q , where D is the number of dimensions and L is the length of the time series.

$$ED(P, Q) = \sum_{j=1}^D \sqrt{\sum_{i=1}^L (p_i^j - q_i^j)^2} \quad (1)$$

A limitation of ED is that it cannot allow for processes that are not directly aligned or which unfold at differing rates.

Distance measures that do make such allowance are known as *elastic distances*. One popular example is the Dynamic Time Warping (DTW) distance. DTW computes the minimum distance of two time series by finding the optimum alignment of two time series and taking into account the temporal order of the data (Lines and Bagnall 2015; Tan et al. 2018, 2020b). Typically, DTW is computed with a warping constraint that limits the warping path (Tan et al. 2018). This has the effect of minimising irregular warping and reducing the time complexity of DTW (Tan et al. 2018, 2020b). Since DTW is a widely studied distance measure, we refer interested readers to the following papers (Tan et al. 2018, 2020b) for more details.

Figure 8a, b illustrate the differences between ED and DTW distance. For multivariate time series, DTW can be computed dependent or independent of the dimensions of the time series, commonly known as DTW_D and DTW_I (Shokoohi-Yekta et al. 2017).

There are various other distance measures other than ED and DTW, none of which dominates the others in terms of classification accuracy, but each of which excels at some tasks (Lines and Bagnall 2015). The Ensembles of Elastic Distances (EE) (Lines and Bagnall 2015) is a combination of 11 elastic distances that is significantly more accurate than each of the individual distances. Although accurate, EE is not computationally efficient as it requires a grid search over a range of parameters for

each elastic distance. FastEE (Tan et al. 2020b) is a significantly faster version of EE that trims the parameter space by leveraging off the properties of each elastic distances. Instead of performing a grid search, Proximity Forest (PF) (Lucas et al. 2019) is a tree-based algorithm where an elastic distance and its parameters are selected at random at each node of the tree. PF has shown to be significantly more accurate and faster than EE for many TSC tasks (Lucas et al. 2019). Although the modification of the NN algorithm for regression tasks is very straightforward, applying EE or PF to regression tasks requires more complex modification of the algorithm which we leave for future work. In this work, we focus only on the two most popular TSC NN algorithms, NN with ED (NN-ED) and DTW distance (NN-DTW).

3.4 Deep learning algorithms

Deep learning models are capable of predicting both discrete labels (classification) and continuous values (regression). Fundamentally, the output of a neural network is a continuous value. Typically for classification tasks, softmax activation is used at the output layer to output class probabilities and classification is done by taking the class with the highest probability. The softmax activation is replaced with linear activation for regression tasks. Apart from the activation functions, the loss function has to be changed as well. The categorical cross entropy loss function that is commonly used for classification can be replaced by either the mean squared error or the mean absolute error loss function for regression tasks, in this case, mean squared error is chosen. Recently, several deep learning models have been developed and benchmarked for TSC (Fawaz et al. 2019; Wang et al. 2017; Fawaz et al. 2018, 2020). In this work, we adapted three TSC deep learning models, Residual Networks (ResNet), Fully Convolutional Neural Networks (FCN) and Inception network (Fawaz et al. 2020).

ResNet and FCN were first proposed in Wang et al. (2017). In a recent survey on deep learning for TSC (Fawaz et al. 2019), ResNet was ranked the most accurate univariate TSC model benchmarked on 85 univariate time series datasets (Dau et al. 2019). ResNet consists of 3 residual blocks with 3 convolutional layers in each block, followed by a global average pooling layer and an output layer. Different from the typical convolutional networks, ResNet has a shortcut residual connection between the convolutional layers which makes training easier by reducing the vanishing gradient effect.

FCN is the most accurate deep learning model for multivariate TSC on 12 multivariate time series datasets (Baydogan and Runger 2015) and the second most accurate deep learning model for univariate TSC. It is composed of three convolutional blocks with batch normalisation and a ReLU activation function. Then, global average pooling is applied to the last convolutional block and connected to a softmax classifier (Fawaz et al. 2019). For regression, the softmax activation function is replaced with linear activation function.

Fawaz et al. (2020) recently proposed the Inception network, which significantly improved existing deep learning models and achieved competitive performance with the state-of-the-art TSC model, HIVE-COTE (Lines et al. 2016). The Inception network consists of two different residual blocks connecting the input to the next block's

input to mitigate the vanishing gradient problem (Fawaz et al. 2020). Each residual block is comprised of three Inception modules. There are two major components in each of the inception module. The first one is the bottleneck layer that reduces the dimension of the time series using m filters and also allowing the Inception network to have ten times longer filters than ResNet (Fawaz et al. 2020). The second component consists of sliding multiple filters of different lengths to the output of the first component. A MaxPooling operation is also applied to the time series in parallel to these two components. The output from each of the convolution and MaxPooling operation is then concatenated to form the output of the Inception module. Finally, global average pooling is applied to the final residual block and passed to a fully connected layer for classification.

In our work, we use the same architecture from the original papers (Fawaz et al. 2019, 2020) with some minor modifications to the activation and loss functions as mentioned above. We refer interested readers to the respective papers for the details of these architectures.

4 Benchmarking results

In this section, we evaluate the regression algorithms described in Sect. 3 and set a baseline using the datasets from our TSER archive (Tan et al. 2020a) described in Sect. 2.2. We evaluate and benchmark the following regression algorithms:

1. FPCR (Goldsmith and Scheipl 2014)
2. FPCR with B-spline (Goldsmith and Scheipl 2014)
3. Grid-search optimised SVR (Drucker et al. 1997)
4. RF (Breiman 2001) with 100 trees
5. XGBoost (Chen and Guestrin 2016) with 100 trees
6. NN-ED with $k = 1, 5$ (1-NN-ED and 5-NN-ED)
7. NN-DTW with $k = 1, 5$ (1-NN-DTW and 5-NN-DTW) and warping window $w = 0.1 \times L$
8. FCN (Fawaz et al. 2019)
9. ResNet (Fawaz et al. 2019)
10. Inception Network (Fawaz et al. 2020)
11. Rocket (Dempster et al. 2020).

Missing values in the time series are linearly interpolated. When using a traditional regression model (i.e. non-temporal), the time series are flattened out into a single long feature vector.

We used the standard Scikit-Learn Python library (Pedregosa et al. 2011) to implement SVR and RF algorithms. The SVR algorithm is optimised by performing a 3-fold cross validation with grid search on the hyper-parameters. Specifically, the kernel, gamma and C parameters are optimised from a standard range of values. The kernel function is selected from RBF and Sigmoid. The gamma parameter selected from [0.001, 0.01, 0.1, 1] defines the influence of support vectors. The regularisation parameter C is selected from [0.1, 1, 10, 100]. XGBoost was implemented using the

Python XGBoost library.³ Apart from the number of trees, we use the default parameters for both RF and XGBoost from the Python libraries. Our empirical experiments show that RF and XGBoost with parameters optimised using a grid search strategy performs similarly or worse than the default parameters and takes a significantly longer time to train. Hence they are excluded from this benchmark. The FPCR and FPCR with B-spline models are implemented using the Scikit-FDA Python package,⁴ a library for functional data analysis in Python.

For time series algorithms, we adapted the code from Fawaz et al. (2019)⁵ for both ResNet and FCN and Fawaz et al. (2020)⁶ for Inception Network. The code for Rocket was taken from Dempster et al. (2020)⁷ and modified for multivariate time series with the help from the original authors. The multivariate version of Rocket applies the transformation to each dimension independently.

The time series NN algorithms were all implemented in Java. Our source code has been made open source online at <https://github.com/ChangWeiTan/TS-Extrinsic-Regression>.

Since some of the algorithms are non-deterministic, we evaluate all the algorithms over 5 runs and report the average root mean squared error (RMSE), one of the most widely used metrics for regression tasks. Equation 2 describes the formal definition of RMSE where n is the number of instances, y_i and \hat{y}_i are the actual and predicted target respectively.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}} \quad (2)$$

We compare the algorithms statistically over the current datasets following the recommendations from (Demšar 2006). First, we rank each algorithm by RMSE for every dataset. Rank 1 is assigned to the algorithm with the lowest RMSE while rank 13 is assigned to the highest one. Fractional ranking is assigned to the algorithm in case of ties. We then compute the average rank for each algorithm. Then, the Friedman test (Friedman 1940; Demšar 2006) was applied to the average ranks. If the null hypothesis is rejected, the post-hoc two-tailed Nemenyi test is used to compare the algorithms to each other (Demšar 2006). Using this test, the performance of the algorithms is significantly different if the average ranks differ by at least the critical difference shown in Eq. 3, where $q_\alpha = 3.313$ is the critical value for $\alpha = 0.05$, $k = 13$ being the number of algorithms and $N = 19$ being the number of datasets. This gives $CD = 4.186$.

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \quad (3)$$

³ https://xgboost.readthedocs.io/en/latest/python/python_intro.html.

⁴ <https://fda.readthedocs.io/en/latest/>.

⁵ <https://github.com/hfawaz/dl-4-tsc>.

⁶ <https://github.com/hfawaz/InceptionTime>.

⁷ <https://github.com/angus924/rocket>.

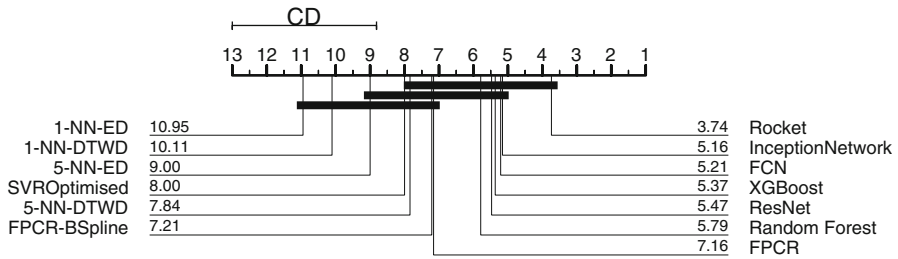


Fig. 9 Critical difference diagram showing statistical difference comparison of 13 regression algorithms on the current regression archive

Finally, a critical difference diagram was used to visualise the comparison, where the thick horizontal line connecting a group of algorithms indicates that all the algorithms in the group are not significantly different from one another (Demšar 2006). Figure 9 shows the critical difference diagram of comparing the algorithms used to benchmark the existing archive. The average ranks are indicated next to the algorithms in the figure.

The critical difference diagram in Fig. 9 shows that Rocket is the most accurate algorithm with an average rank of 3.74 and is significantly different from 1-NN-ED and 1-NN-DTWD. The figure also shows that there is no significant difference between the state-of-the-art time series algorithms and classical regression algorithms. However, our experiments indicate that Rocket is the most computationally efficient compared to all other algorithms.

We further compare the relative performance of each algorithm on the current TSER archive. The relative performance of an algorithm is computed by scaling the RMSE of each algorithm with the median RMSE obtained for a dataset. Equation 4 describes the equation to scale the RMSE of algorithm j for dataset i .

$$scaled_RMSE_i^j = \frac{RMSE_i^j}{RMSE_i^j + median(RMSE_i)} \tag{4}$$

The algorithm with median RMSE can be interpreted as the algorithm that gives the average performance for the dataset. Hence, values larger than 0.5 indicate worse performance, while values smaller than 0.5 indicate a better performance than the average performance. Figure 10 illustrates the scaled RMSE for each algorithm in the form of boxplots. It shows that most algorithms have their values around 0.5, while bespoke time series algorithms such as Rocket, FCN, ResNet and Inception Network have larger spread in the values and tend to have values smaller than 0.5. This implies that when time series algorithms perform better, they perform significantly better than the other algorithms, while the other algorithms tend to perform similarly to an average algorithm. The median of all algorithms are similar, around 0.5, which suggests that there is room for better algorithms to be developed for TSER problems.

Table 2 shows a more detailed results of these algorithms on all the datasets in the archive. The results show that Rocket performs the best overall with the lowest

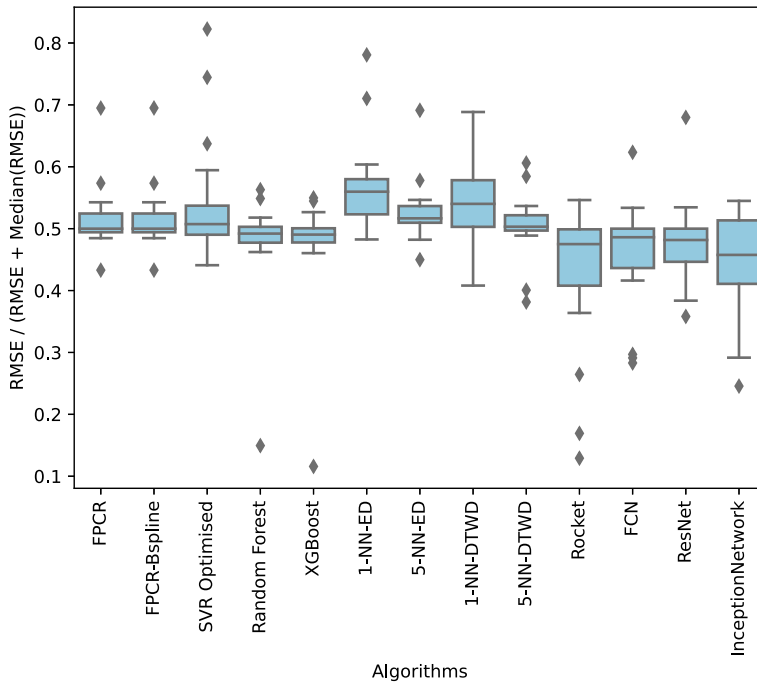


Fig. 10 The relative RMSE of each algorithm on the current TSER archive. Values greater than 0.5 indicate that the algorithm has RMSE higher than the average algorithm while values smaller than 0.5 indicate an RMSE lower than the average algorithm

average RMSE ranks followed by the other state-of-the-art TSC algorithms. RF and XGBoost are both very competitive compared with the time series algorithms. This is expected as XGBoost and RF are both the state of the art in ML algorithms, especially in popular data science and ML competitions (Nielsen 2016). The results also indicate that the SoFR algorithms are also competitive as they are not significantly different from the standard regression algorithms. This further strengthens our findings from Fig. 10 that there is room for better algorithms to be developed for TSER problems and that new algorithms should also be computationally efficient.

On the tasks of energy and health monitoring, time series algorithms are clearly performing better than classical regression algorithms, with the top 3 algorithms being time series algorithms. For instance, Inception network performs the best on heart rate prediction tasks while Rocket is the most accurate on energy prediction tasks. There is no clear winner for environment monitoring tasks. Classical regression algorithms perform better at predicting pollution level while time series algorithms perform better on the remaining datasets. The reason is that, the pollution metrics from these pollution datasets can be estimated fairly easily by applying a threshold to the measurements from gas sensors, where classical regression algorithms such as RF and XGBoost are very good at. Nonetheless, we expect a TSER algorithm that uses feature extraction techniques such as the TSC counterparts, Shapelet Transform (Lines et al. 2012), Time

Series Forest (Deng et al. 2013) and BOSS (Schäfer 2015), will perform better than classical regression algorithms.

Although there is also no clear winner on the new sentiment analysis task that we propose in this work, the results show that predicting sentiment scores using time series data is feasible, achieving very low RMSE scores. Both classical regression and time series algorithms perform similarly on forecasting tasks. This is expected as both types of algorithms are not designed for forecasting and we expect that a forecasting algorithm if adapted for TSER will perform better. Besides, the small Covid3Month dataset with 140 time series of length 84 may not have enough data for the algorithms to train on. Overall, the results indicate that there is a need to design better TSER algorithms that can better generalise for most datasets.

5 Conclusion and future work

In this paper, we introduced and motivated the *Time Series Extrinsic Regression* problem where the goal is to predict a continuous value using time series data. We showed some examples of real-life applications where TSER may be useful and discussed some existing methods for this task. We benchmarked these methods on the first TSER benchmarking archive and showed that Rocket, one of the state-of-the-art TSC algorithms performs the best overall. Although time series specific Rocket achieved the highest overall rank on accuracy, its rank is not statistically distinguishable from classical machine learning algorithms XGBoost and Random Forest that ignore the temporal order of the data. This is in contrast to the state-of-the-art in time series classification, where bespoke algorithms significantly outperform approaches that ignore the temporal information in the data. Therefore, this suggests much research is needed to develop better algorithms to improve accuracy on TSER problems.

Table 2 RMSE obtained for the different algorithms on the TSER archive

Dataset	RMSE												
	FPCR	FPCR-BSpline	SVR Optimised	Random Forest	XGBoost	1-NN-ED	5-NN-ED	1-NN-DTWD	5-NN-DTWD	Rocket FCN	ResNet Inception		
AppliancesEnergy	5.41	5.41	3.46	3.46	3.49	5.23	4.23	6.04	4.02	2.30	2.87	3.07	4.44
HouseholdPower	147.55	147.55	152.39	248.86	231.09	473.93	432.60	427.04	297.22	132.80	162.24	193.21	153.72
Consumption1	46.93	46.93	55.98	46.93	44.37	71.48	64.27	58.80	51.50	32.61	46.83	39.08	39.41
Consumption2	11.09	11.10	4.79	0.86	0.64	6.54	5.85	4.98	4.87	3.36	4.99	4.06	1.59
Benzene Concentration	99.73	99.73	110.57	94.07	93.14	139.23	115.67	139.14	115.50	120.06	94.35	95.49	96.75
BeijingPM10Quality	69.38	69.37	75.73	63.30	59.50	88.19	74.16	88.26	72.72	62.77	59.73	64.46	62.23
BeijingPM25Quality	37.68	37.68	39.73	32.16	32.44	47.84	38.54	39.97	35.19	29.41	33.26	30.35	28.80
LiveFuelMoisture Content	0.02	0.02	0.05	0.02	0.02	0.02	0.02	0.01	0.01	0.00	0.01	0.01	0.02
FloodModeling1	0.02	0.02	0.08	0.01	0.02	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.01
FloodModeling2	0.02	0.02	0.04	0.02	0.02	0.02	0.02	0.01	0.01	0.00	0.01	0.02	0.01
FloodModeling3	0.02	0.02	0.04	0.02	0.02	0.02	0.02	0.01	0.01	0.00	0.01	0.02	0.01

Table 2 continued

Dataset	RMSE		FPCR-BSpline	SVR Optimised	Random Forest	XGBoost	1-NN-ED	5-NN-ED	1-NN-DTWD	5-NN-DTWD	Rocket	FCN	ResNet	Inception
	FPCR	B-Spline												
Australia Rainfall	8.44	8.44	8.65	8.39	8.49	30.25	10.23	12.00	11.95	8.12	8.43	8.18	8.84	
PPGDalia	20.67	20.67	19.01	17.53	16.58	21.88	18.28	26.03	20.77	14.05	13.04	11.38	9.92	
IEEEPPG	31.38	31.38	37.25	32.11	31.49	33.21	27.11	37.14	33.57	36.52	34.33	33.15	23.90	
BIDMC32HR	13.98	13.98	13.39	15.02	13.96	14.84	14.76	15.29	15.13	13.94	13.13	10.74	9.43	
BIDMC32RR	3.37	3.37	3.17	4.35	4.37	4.39	4.14	3.53	3.43	4.09	3.58	3.92	3.02	
BIDMC32SpO2	4.95	4.95	4.80	4.57	4.45	5.53	5.41	5.22	5.12	5.22	5.97	5.99	5.58	
NewsHeadline Sentiment	0.14	0.14	0.14	0.15	0.14	0.20	0.16	0.20	0.16	0.14	0.15	0.15	0.15	
NewsTitle Sentiment	0.14	0.14	0.14	0.14	0.14	0.19	0.15	0.19	0.15	0.14	0.14	0.14	0.16	
Covid3Month	0.05	0.05	0.07	0.04	0.05	0.05	0.04	0.05	0.04	0.04	0.07	0.10	0.05	
Average rank	7.16	7.21	8.00	5.79	5.37	10.95	9.00	10.11	7.84	3.74	5.21	5.47	5.16	

The lowest RMSE per dataset is indicated in bold

Acknowledgements This research has been supported by Australian Research Council grant DP210100072; and the Air Force Office of Scientific Research, Asian Office of Aerospace Research and Development (AOARD) under award number FA2386-18-1-4030. The authors appreciate the data donation from all the donors and would like to thank the authors of Fawaz et al. (2019) and Dempster et al. (2020) for providing their source code online.

References

- Bagnall A, Lines J, Hills J, Bostrom A (2015) Time-series classification with COTE: the collective of transformation-based ensembles. *IEEE Trans Knowl Data Eng* 27(9):2522–2535
- 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 Min Knowl Discov* 31(3):606–660
- Baydogan MG, Runger G (2015) Learning a symbolic representation for multivariate time series classification. *Data Min Knowl Discov* 29(2):400–422
- Box GE, Jenkins GM (1970) Time series analysis forecasting and control. Tech. rep., Wisconsin University, Dept of Statistics
- Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
- Chatfield C (1978) The Holt-Winters forecasting procedure. *J R Stat Soc Ser C (Appl Stat)* 27(3):264–279
- Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 785–794
- Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn* 20(3):273–297
- Dau HA, Bagnall A, Kamgar K, Yeh CCM, Zhu Y, Gharghabi S, Ratanamahatana CA, Keogh E (2019) The UCR time series archive. *IEEE/CAA J Autom Sin* 6(6):1293–1305
- De Vito S, Massera E, Piga M, Martinotto L, Di Francia G (2008) On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sens Actuators B Chem* 129(2):750–757
- Dempster A, Petitjean F, Webb GI (2020) ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Min Knowl Discov* 34(5):1454–1495
- Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res* 7:1–30
- Deng H, Runger G, Tuv E, Vladimir M (2013) A time series forest for classification and feature extraction. *Inf Sci* 239:142–153
- Drucker H, Burges CJ, Kaufman L, Smola AJ, Vapnik V (1997) Support vector regression machines. In: *Advances in neural information processing systems*, pp 155–161
- Dua D, Graff C (2017) UCI machine learning repository. <http://archive.ics.uci.edu/ml>
- Fawaz HI, Forestier G, Weber J, Idoumghar L, Muller PA (2018) Transfer learning for time series classification. In: Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), pp 1367–1376
- Fawaz HI, Forestier G, Weber J, Idoumghar L, Muller PA (2019) Deep learning for time series classification: a review. *Data Min Knowl Discov* 33(4):917–963
- Fawaz HI, Lucas B, Forestier G, Pelletier C, Schmidt DF, Weber J, Webb GI, Idoumghar L, Muller PA, Petitjean F (2020) Inceptiontime: finding alexnet for time series classification. *Data Min Knowl Discov* 34(6):1936–1962
- Friedman M (1940) A comparison of alternative tests of significance for the problem of m rankings. *Ann Math Stat* 11(1):86–92
- Fulcher BD, Little MA, Jones NS (2013) Highly comparative time-series analysis: the empirical structure of time series and their methods. *J R Soc Interface* 10(83):20130048. <https://doi.org/10.1098/rsif.2013.0048>
- Gardner ES Jr (1985) Exponential smoothing: the state of the art. *J Forecast* 4(1):1–28
- Goldsmith J, Scheipl F (2014) Estimator selection and combination in scalar-on-function regression. *Comput Stat Data Anal* 70:362–372
- Grabocka J, Schilling N, Wistuba M, Schmidt-Thieme L (2014) Learning time-series shapelets. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, pp 392–401
- Hyndman R (2018) A brief history of time series forecasting competitions

- Hyndman R, Koehler AB, Ord JK, Snyder RD (2008) Forecasting with exponential smoothing: the state space approach. Springer, Berlin
- Kang Y, Hyndman RJ, Smith-Miles K (2017) Visualising forecasting algorithm performance using time series instance spaces. *Int J Forecast* 33(2):345–358. <https://doi.org/10.1016/j.ijforecast.2016.09.004>
- Karlen W, Turner M, Cooke E, Dumont G, Ansermino JM (2010) Capnobase: signal database and tools to collect, share and annotate respiratory signals. In: Annual meeting of the Society for Technology in Anesthesia (STA), West Palm Beach, p 25
- Lin J, Khade R, Li Y (2012) Rotation-invariant similarity in time series using bag-of-patterns representation. *J Intell Inf Syst* 39(2):287–315
- Lines J, Bagnall A (2015) Time series classification with ensembles of elastic distance measures. *Data Min Knowl Discov* 29(3):565–592
- Lines J, Davis LM, Hills J, Bagnall A (2012) A shapelet transform for time series classification. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining, pp 289–297
- Lines J, Taylor S, Bagnall A (2016) HIVE-COTE: the hierarchical vote collective of transformation-based ensembles for time series classification. In: Proceedings of the 16th IEEE International Conference on Data Mining (ICDM), pp 1041–1046
- Lubba CH, Sethi SS, Knaute P, Schultz SR, Fulcher BD, Jones NS (2019) catch22: canonical time-series characteristics. *Data Min Knowl Discov* 33(6):1821–1852. <https://doi.org/10.1007/s10618-019-00647-x>
- Lucas B, Shifaz A, Pelletier C, O’Neill L, Zaidi N, Goethals B, Petitjean F, Webb GI (2019) Proximity forest: an effective and scalable distance-based classifier for time series. *Data Min Knowl Discov* 33(3):607–635
- Makridakis S, Hibon M (2000) The M3-competition: results, conclusions and implications. *Int J Forecast* 16(4):451–476
- Makridakis S, Andersen A, Carbone R, Fildes R, Hibon M, Lewandowski R, Newton J, Parzen E, Winkler R (1982) The accuracy of extrapolation (time series) methods: results of a forecasting competition. *J Forecast* 1(2):111–153
- Makridakis S, Spiliotis E, Assimakopoulos V (2018) The M4 competition: results, findings, conclusion and way forward. *Int J Forecast* 34(4):802–808
- Makridakis S, Spiliotis E, Assimakopoulos V (2020) The M4 competition: 100,000 time series and 61 forecasting methods. *Int J Forecast* 36(1):54–74
- Meredith DJ, Clifton D, Charlton P, Brooks J, Pugh C, Tarassenko L (2012) Photoplethysmographic derivation of respiratory rate: a review of relevant physiology. *J Med Eng Technol* 36(1):1–7
- Moniz N, Torgo L (2018) Multi-source social feedback of online news feeds. *arXiv preprint arXiv:1801.07055*
- Montero-Manso P, Athanasopoulos G, Hyndman RJ, Talagala TS (2020) Fforma: feature-based forecast model averaging. *Int J Forecast* 36(1):86–92. <https://doi.org/10.1016/j.ijforecast.2019.02.011>
- Mueen A, Keogh E, Young N (2011) Logical-shapelets: an expressive primitive for time series classification. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, pp 1154–1162
- Nielsen D (2016) Tree boosting with xgboost-why does xgboost win every machine learning competition? Master’s thesis, NTNU
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. *J Mach Learn Res* 12:2825–2830
- Pelletier C, Webb GI, Petitjean F (2019) Temporal convolutional neural network for the classification of satellite image time series. *Remote Sens* 11(5):523
- Pimentel MA, Charlton PH, Clifton DA (2015) Probabilistic estimation of respiratory rate from wearable sensors. In: *Wearable electronics sensors*. Springer, pp 241–262
- Pimentel MA, Johnson AE, Charlton PH, Birrenkott D, Watkinson PJ, Tarassenko L, Clifton DA (2016) Toward a robust estimation of respiratory rate from pulse oximeters. *IEEE Trans Biomed Eng* 64(8):1914–1923
- Rakthanmanon T, Keogh E (2013) Fast shapelets: a scalable algorithm for discovering time series shapelets. In: Proceedings of the 2013 SIAM International Conference on Data Mining (SDM). SIAM, pp 668–676

- Reiss PT, Goldsmith J, Shang HL, Ogden RT (2017) Methods for scalar-on-function regression. *Int Stat Rev* 85(2):228–249
- Reiss A, Indlekofer I, Schmidt P, Van Laerhoven K (2019) Deep PPG: large-scale heart rate estimation with convolutional neural networks. *Sensors* 19(14):3079
- Salehizadeh S, Dao D, Bolkhovskiy J, Cho C, Mendelson Y, Chon KH (2016) A novel time-varying spectral filtering algorithm for reconstruction of motion artifact corrupted heart rate signals during intense physical activities using a wearable photoplethysmogram sensor. *Sensors* 16(1):10
- Sammut C, Webb GI (2011) *Encyclopedia of machine learning*. Springer, Berlin
- Schäcker T, Muma M, Zoubir AM (2017) Computationally efficient heart rate estimation during physical exercise using photoplethysmographic signals. In: 2017 25th European Signal Processing Conference (EUSIPCO). IEEE, pp 2478–2481
- Schäfer P (2015) The BOSS is concerned with time series classification in the presence of noise. *Data Min Knowl Discov* 29(6):1505–1530
- Schäfer P, Leser U (2017a) Fast and accurate time series classification with weasel. In: Proceedings of the 2017 ACM on conference on information and knowledge management, pp 637–646
- Schäfer P, Leser U (2017b) Multivariate time series classification with WEASEL+MUSE. arXiv preprint [arXiv:1711.11343](https://arxiv.org/abs/1711.11343)
- Segal MR (2004) *Machine learning benchmarks and random forest regression*. UCSF: Center for Bioinformatics and Molecular Biostatistics
- Senin P, Malinchik S (2013) SAX-VSM: interpretable time series classification using SAX and vector space model. In: 2013 IEEE 13th international conference on data mining. IEEE, pp 1175–1180
- Shokoohi-Yekta M, Hu B, Jin H, Wang J, Keogh E (2017) Generalizing DTW to the multi-dimensional case requires an adaptive approach. *Data Min Knowl Discov* 31(1):1–31
- Tan CW, Herrmann M, Forestier G, Webb GI, Petitjean F (2018) Efficient search of the best warping window for dynamic time warping. In: Proceedings of the 2018 SIAM International Conference on Data Mining (SDM). SIAM, pp 225–233
- Tan CW, Bergmeir C, Petitjean F, Webb GI (2020a) Monash University, UEA, UCR time series extrinsic regression archive. arXiv preprint [arXiv:2006.10996](https://arxiv.org/abs/2006.10996)
- Tan CW, Petitjean F, Webb GI (2020b) FastEE: fast ensembles of elastic distances for time series classification. *Data Min Knowl Discov* 34(1):231–272
- Wang Z, Yan W, Oates T (2017) Time series classification from scratch with deep neural networks: a strong baseline. In: 2017 International Joint Conference on Neural Networks (IJCNN). IEEE, pp 1578–1585
- Ye L, Keogh E (2009) Time series shapelets: a new primitive for data mining. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data Mining, pp 947–956
- Yebra M, Quan X, Riaño D, Larraondo PR, van Dijk AI, Cary GJ (2018) A fuel moisture content and flammability monitoring methodology for continental Australia based on optical remote sensing. *Remote Sens Environ* 212:260–272
- Zhang Z (2015) Photoplethysmography-based heart rate monitoring in physical activities via joint sparse spectrum reconstruction. *IEEE Trans Biomed Eng* 62(8):1902–1910
- Zhang Z, Pi Z, Liu B (2014) Troika: a general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise. *IEEE Trans Biomed Eng* 62(2):522–531

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.