sensors

Review

A Systematic Review of Time Series Classification Techniques Used in Biomedical Applications

Will Ke Wang, Ina Chen, Leeor Hershkovich, Jiamu Yang, Ayush Shetty, Geetika Singh, Yihang Jiang, Aditya Kotla, Jason Zisheng Shang, Rushil Yerrabelli, Ali R. Roghanizad, Md Mobashir Hasan Shandhi and Jessilyn Dunn *

Biomedical Engineering Department, Duke University, Durham, NC 27708, USA
* Correspondence: jessilyn.dunn@duke.edu

Abstract: *Background:* Digital clinical measures collected via various digital sensing technologies such as smartphones, smartwatches, wearables, and ingestible and implantable sensors are increasingly used by individuals and clinicians to capture the health outcomes or behavioral and physiological characteristics of individuals. Time series classification (TSC) is very commonly used for modeling digital clinical measures. While deep learning models for TSC are very common and powerful, there exist some fundamental challenges. This review presents the non-deep learning models that are commonly used for time series classification in biomedical applications that can achieve high performance. *Objective:* We performed a systematic review to characterize the techniques that are used in time series classification of digital clinical measures throughout all the stages of data processing and model building. *Methods:* We conducted a literature search on PubMed, as well as the Institute of Electrical and Electronics Engineers (IEEE), Web of Science, and SCOPUS databases using a range of search terms to retrieve peer-reviewed articles that report on the academic research about digital clinical measures from a five-year period between June 2016 and June 2021. We identified and categorized the research studies based on the types of classification algorithms and sensor input types. *Results:* We found 452 papers in total from four different databases: PubMed, IEEE, Web of Science Database, and SCOPUS. After removing duplicates and irrelevant papers, 135 articles remained for detailed review and data extraction. Among these, engineered features using time series methods that were subsequently fed into widely used machine learning classifiers were the most commonly used technique, and also most frequently achieved the best performance metrics (77 out of 135 articles). Statistical modeling (24 out of 135 articles) algorithms were the second most common and also the second-best classification technique. *Conclusions:* In this review paper, summaries of the time series classification models and interpretation methods for biomedical applications are summarized and categorized. While high time series classification performance has been achieved in digital clinical, physiological, or biomedical measures, no standard benchmark datasets, modeling methods, or reporting methodology exist. There is no single widely used method for time series model development or feature interpretation, however many different methods have proven successful.

Keywords: systematic review; time series classification; digital clinical measures; machine learning; feature engineering

check for updates

Citation: Wang, W.K.; Chen, I.; Hershkovich, L.; Yang, J.; Shetty, A.; Singh, G.; Jiang, Y.; Kotla, A.; Shang, J.Z.; Yerrabelli, R.; et al. A Systematic Review of Time Series Classification Techniques Used in Biomedical Applications. *Sensors* **2022**, *22*, 8016. https://doi.org/10.3390/s22208016

Academic Editor: Enrico G. Caiani

Received: 16 August 2022
Accepted: 17 October 2022
Published: 20 October 2022

1. Introduction

Time Series Classification (TSC) involves building predictive models that output a target variable or label from inputs of longitudinal or sequential observations across some time period [1]. These inputs could be from a single variable measured across time or multiple variables measured across time, where the measurements can be ordinal or numerical (discrete or continuous).

Time series data are a very common form of data, containing information about the (changing) state of any variable. Some common examples include stock market prices and temperature values across some period of time. Time series modeling tasks include classification, regression, and forecasting. There are unique challenges that come with modeling time series, given that measurements in time obtained in real-life settings are subject to random noise, and that any measurement at a particular point in time could be related to or influenced by measurements at other points in time [1]. Given this nature of time series data, it is impractical to simply utilize established machine learning algorithms such as logistic regression, support vector machine, or random forest on the raw time series datasets because these data violate the basic assumptions of those models. In recent years, two vastly different camps of time series classification techniques have emerged: deep-learning-based models vs non-deep-learning-based models. While deep learning models are extremely powerful and show great promise in classification performance and generalizability, they also present challenges in the areas of hyperparameter tuning, training, and model complexity decisions. To enable the evaluation of new models, a reasonable baseline is also needed for comparison. Further, there already exists a review on deep-learning time series classification methods [2]. Therefore, the focus of this review is on non-deep learning-based time series classification models.

This paper also focuses specifically on the biomedical applications of time series classification because there has been a huge increase in the generation of biomedical time series datasets (such as data from wearable devices like Apple Watch and Fitbit) recently as well as research using such data—examples include electrocardiogram (ECG, for cardiovascular dysfunction screening) [3], electroencephalogram (EEG, for brainwave tracking) [4], accelerometry (for activity recognition), and polysomnography (PSG for sleep tracking) [5], etc. In addition, an increasing number of people use smart devices or wearables regularly [6] for general fitness tracking [7], sleep tracking [8], fall detection [9], or arrhythmia detection [10]. There is a growing need to design better data mining and classification methods to discern important and useful information from biomedical time series data. This would lead to more reliable methods for screening, diagnosis, and monitoring, thereby providing huge benefits for healthcare as a whole.

Biomedical time series data collected from human subjects often present challenges that impede the ability to leverage time series modeling techniques that are common in other fields. For example, biomedical time series datasets often include just a small number of human subjects due to the resources and effort needed for data collection and annotation (or labeling to produce ground truth), which makes applying deep learning models very difficult since they are extremely data hungry [11]. Another challenge is the non-ergodic nature of datasets collected from human subjects, meaning that human subjects have vast individual differences in mental and physical states, and thereby producing data that look very different from one subject to another [12]. This results in sample level observations or models that perform well on some individuals even while being completely useless for others.

While both reviews and experimental evaluations of recent algorithmic advances have been done [13], the usefulness and applicability of machine learning algorithms is also impacted by interpretability and simplicity, particularly for biomedical predictive or diagnostic tasks. This review systematically surveys papers published in recent years that have used time series classification machine learning algorithms on biomedical datasets to answer the following questions:

(1) What are the most common time series classification algorithms used in biomedical data science in the past six years?
(2) What are the best performing time series classification algorithms for common biomedical signals?
(3) How is interpretability addressed in the scientific literature that describes applying TSC algorithms for specific biomedical tasks?

The motivation for this review came from the observation that the types of algorithms explored and the depth of analysis performed in time series biomedical data science have not been well described. In general, there has been a strong emphasis on algorithmic performance and a lack of focus on interpretability and model simplicity. This review aims to provide a general and recent landscape of the types of time series classification algorithms on longitudinal biomedical data and these algorithms can be applied toward specific tasks and with more insightful analysis.

2. Methods

A list of search terms was developed that are specific for each of the following four databases: PubMed, IEEE, SCOPUS, and Web of Science (Supplementary Table S1). Four literature databases were searched: PubMed, IEEE, Web of Science, and SCOPUS. Among these databases, IEEE enforces a limit on the number of search terms to a maximum of 20. Hence, the defined search terms on IEEE were different from those on PubMed, Web of Science, and SCOPUS. The literature search was limited to the last six years for a manageable scope of review. The defined searches include the general terms and variations of time series machine learning classification and the fields of biomedical data. While the approach limits the coverage of the review, more recent work is often built upon previous work and new time series classification techniques are often compared to established techniques from previous work, therefore this method is expected to provide a sufficient representation of the field. Covidence was used for literature screening and data extraction. This review has a very clear focus on only non-deep-learning time series classification techniques utilized on biomedical data. This boundary notably excludes time series regression tasks and deep learning techniques.

There were two phases of screening before data extraction. The first phase was screening by titles and abstracts, which Covidence automatically extracted from the DOI URLs. This phase was completed by two reviewers where each reviewer read through each title/abstract and labeled them as “include” or “exclude”. Conflicts were resolved by discussion among the reviewers in order to reach consensus. A total of 260 papers were found to be irrelevant in this phase, mainly due to the following reasons:

- Classification algorithm is not used on biomedical data or time series data;
- The article does not focus on classification algorithms, but regression algorithms, clustering algorithms, or other algorithms;
- The article focuses only on deep learning algorithms.

The second phase was screening of the full texts, which were pulled automatically by Covidence if free full texts were readily available. The rest of the full texts were uploaded manually to the Covidence platform using university credentials for access. Again, two reviewers each went through all the papers and adjudicated the inclusion of papers into the final data extraction. Conflicts were resolved by discussion to reach a consensus. A total of 40 papers were excluded in this phase, mainly due to the following reasons:

- No access to full paper;
- Not enough information about classification algorithm performance included;
- The data came from animals instead of humans;
- The algorithms used are not classification algorithms.

Data from each paper were extracted by one of five reviewers, and then verified, edited, and cleaned by the study lead. In Table 1, we detail the information that was extracted from each paper:

Table 1. Data fields extracted from identified academic research.

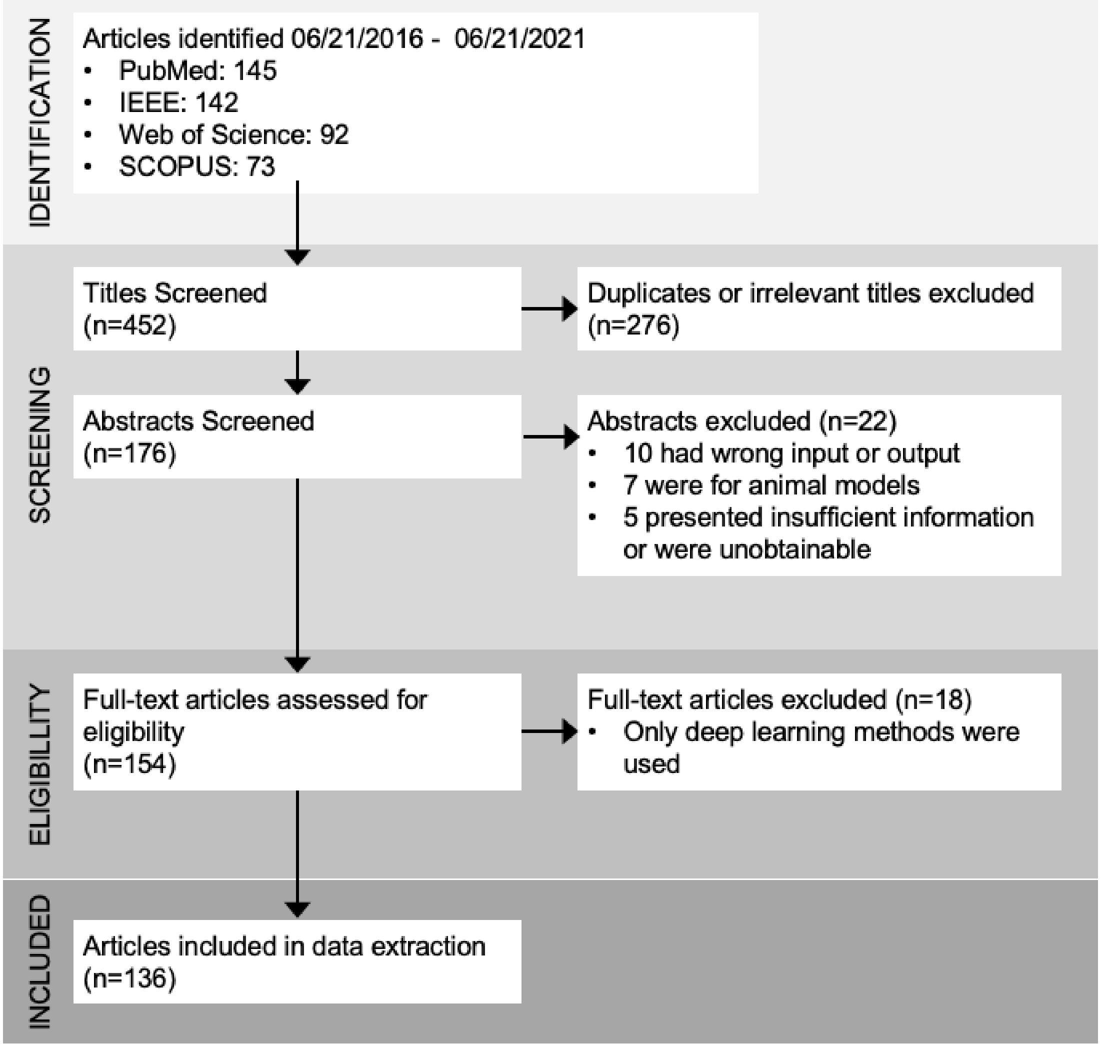
Categories	Choices/Sub-Fields	Definitions/Descriptions
General (Relevant) Information	Article Type	The type of article for a particular paper being reviewed, such as Journal Article/Conference Article/Review Paper
	Area of application	Describes the area of biomedical signal and application this paper is about.
	Aim of study	Defines the specific challenge or question this paper is aimed at tackling
	Name of Publisher/Journal/Conference	Site of article publication.
	Classification Task	Defines the kind of classification task performed in this article. (Pointwise classification, window classification, or whole sequence classification)
	Input data (X)	The type of input biomedical time series data
	Label (Y)	The output label or variable. Example: sleep vs wake, healthy vs diseased.
	Data source or open dataset name	States if the data are open source and where the dataset is hosted.
	Population Size	The number of subjects are included in this dataset.
	Data exclusion criteria	States the criteria considered to exclude subjects or specific parts of the data.
	All algorithms tested	List (or examples) of all the algorithms tested.
	Best algorithm name	The name of the best algorithm.
Classification Task	Whole-Series Classification	In whole time series classification (WSC) for a dataset of n samples, we are provided a set of tuples where each of an entire time series is associated with one class label.
	Sequence-to-sequence (point-wise)	The class label of each point in time is predicted.
	Window-based Classification or Onset Detection	Onset detection is a subtype of time series classification in which—as opposed to whole series classification—class labels are provided with a time-stamp. As an alternative to time pointwise classification, time-stamped labels have been leveraged for classifying time series windows that precede the class label's time-stamp. For onset detection, a class label requires a time-stamp. This additional information can enforce that solely information from the past and present is used to predict a future target. This can be understood as a compromise between time pointwise classification and whole time series classification. An example is to detect the onset of sepsis in the intensive care unit
Best Algorithm Class	Feature Engineering	The type of time series classification technique where features are extracted to describe a particular time series sample and the features are fed into traditional machine learning algorithms as inputs of the predictive modeling.
	Statistical Modeling	This technique uses statistical modeling (such as Kalman filters or state-space models like Hidden Markov Models) to describe or fit the time series observed. Using the information obtained from statistical models, we can make decisions or extract features to be used as inputs to machine learning algorithms.
	Wavelet Transform [8]	Wavelet Transform can be used for signal cleaning (preprocessing), signal decomposition (preprocessing), and feature extraction. This technique is widely used and can be considered an integral part of time series machine learning.

Table 1. *Cont.*

Categories	Choices/Sub-Fields	Definitions/Descriptions
	Distance-based methods [7]	This method is based on defining or quantifying the difference or distance (proxy for dissimilarity) between every pair of time series data samples in the dataset. Classification is performed based on the calculated distances, where two time series that are in close proximity (i.e., they have a small distance) under some distance measure are likely to come from the same class.
	Ensemble-based	Ensemble-based classification algorithms utilize multiple algorithms to make predictions and then aggregate the results coming from these different algorithms
	Shapelet/Shape-based	Shapelet-based methods are similar to significant pattern mining. Time series shapelets are subsequences that maximize classification performance.
	Non-linear index and thresholding	This time series classification method is based on defining indices based on domain- and data-driven time series features. The thresholds for these indices can be predefined or found through statistical learning. The thresholds are then used to make predictions of classes.
	Other	Any other methods of time series classification that cannot be easily categorized.
Best Algorithm Performances [14,15]	Accuracy	The degree of correctness of a calculation of the best algorithm reported. $\frac{TP+TN}{TP+FN+TN+FP}$
	F1-score	The harmonic mean of precision and recall of the best algorithm reported. $2 \times \frac{Precision \times Sensitivity}{Precision+Sensitivity}$
	Area Under Curve of Receiver-Operating Characteristic	The measure of the usefulness of a test, in general, of the best algorithm reported.
	Sensitivity	The percentage of true positives of the best algorithm reported. $\frac{TP}{TP+FN}$
	Specificity	The percentage of true negatives of the best algorithm reported. $\frac{TN}{TN+FP}$
	Cohen's Kappa	A statistical measure of inter-rater reliability for categorical variables of the best algorithm reported. $\frac{p_0-p_e}{1-p_e}$
	Positive Predictive Value	The percentage of positive test results is a true positive. $\frac{TP}{TP+FP}$
	Negative Predictive Value	The percentage of negative test results is a true negative. $\frac{TN}{TN+FN}$
	False Positive Rate	The percentage of false alarm of the best algorithm reported $\frac{FP}{FP+TN}$
	Area Under Precision-Recall Curve	A model performance metric for binary responses that is appropriate for rare events and not dependent on model specificity

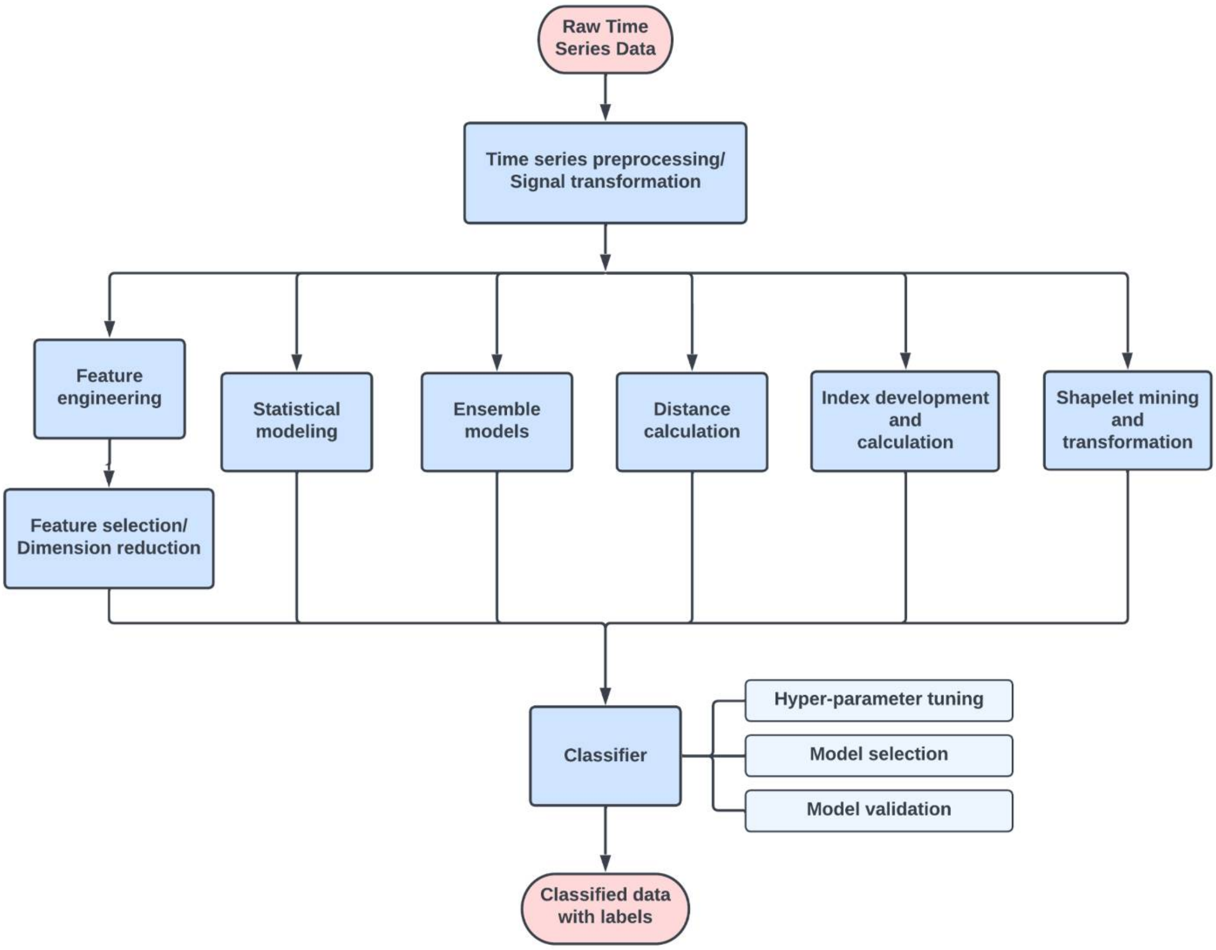
3. Results

After removing duplicates and irrelevant papers, 135 articles remained for review and data extraction. Time series classification modeling typically consists of 3 main steps: signal preprocessing/transformation, modeling, and classification (Figure 1). The classification step is basically the process of model tuning, training, and validation. The different types of algorithms used in the modeling steps are adapted from the categories summarized in "The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances" by Ruiz et al. [13] (Figure 1b). The most common techniques found in our search are feature engineering and selection, statistical modeling, distance-based, index development, and shape-based methods (Table 1 and Table S2).


IDENTIFICATION
Articles identified 06/21/2016 - 06/21/2021
• PubMed: 145
• IEEE: 142
• Web of Science: 92
• SCOPUS: 73
SCREENING
Titles Screened (n=452)
Duplicates or irrelevant titles excluded (n=276)
Abstracts Screened (n=176)
Abstracts excluded (n=22)
• 10 had wrong input or output
• 7 were for animal models
• 5 presented insufficient information or were unobtainable
ELIGIBILLITY
Full-text articles assessed for eligibility (n=154)
Full-text articles excluded (n=18)
• Only deep learning methods were used
INCLUDED
Articles included in data extraction (n=136)

(**a**)

Figure 1. *Cont.*

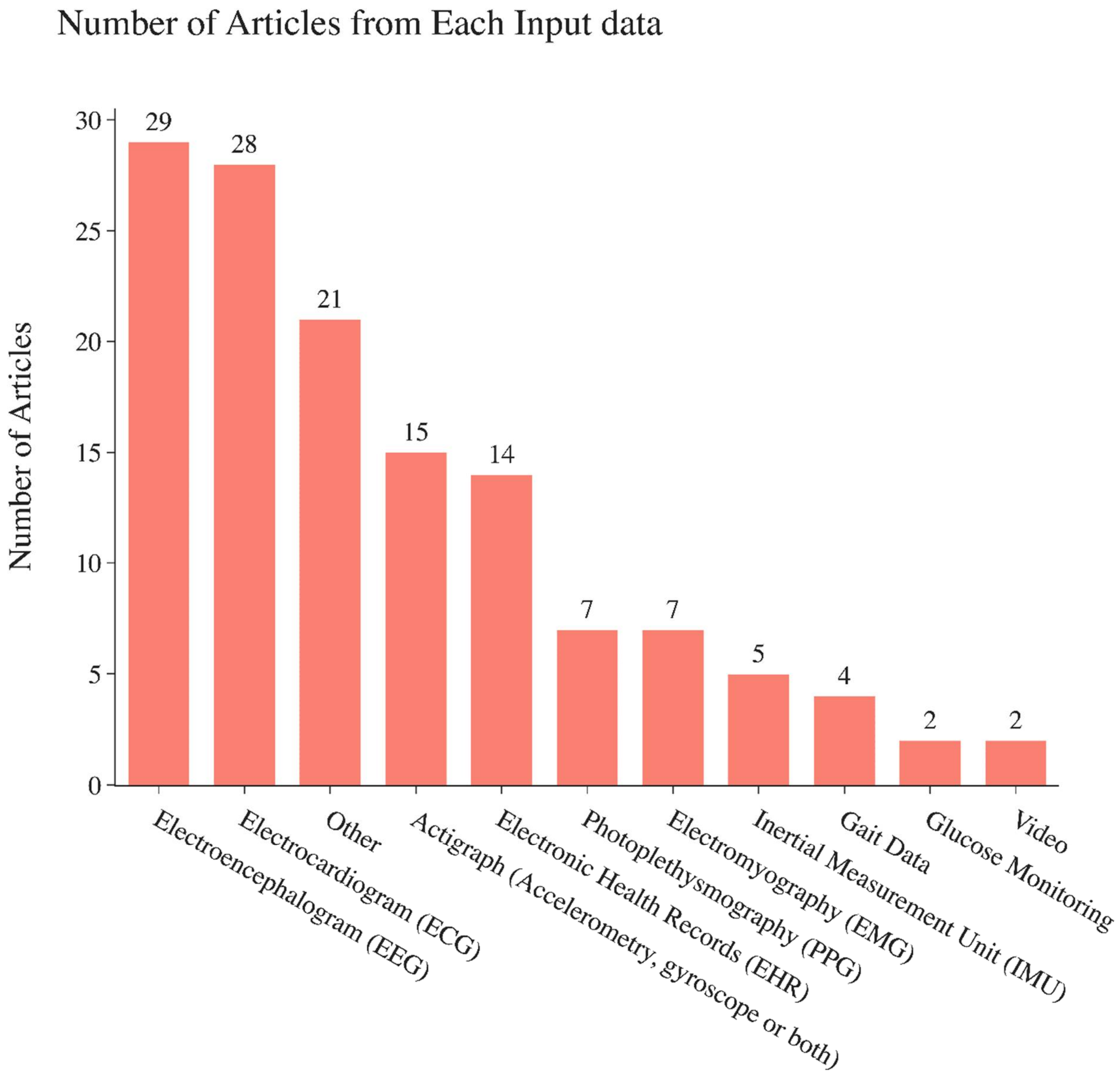

Raw Time Series Data
Time series preprocessing/ Signal transformation
Feature engineering
Feature selection/ Dimension reduction
Statistical modeling
Ensemble models
Distance calculation
Index development and calculation
Shapelet mining and transformation
Classifier
Hyper-parameter tuning
Model selection
Model validation
Classified data with labels

(**b**)

Figure 1. (**a**). Review results and the number of papers through each selection process. (**b**). Flow chart of the common steps in time series classification techniques found in this review. Raw time series signals usually go through some steps of preprocessing for artifact removal or noise reduction, and then are passed through the modeling stage. The modeling stage can use many different types of algorithms, such as feature engineering and selection, statistical modeling, and distance calculation (Table 1). Classifiers are then tuned, trained, validated, and compared to find the best model for a specific task.

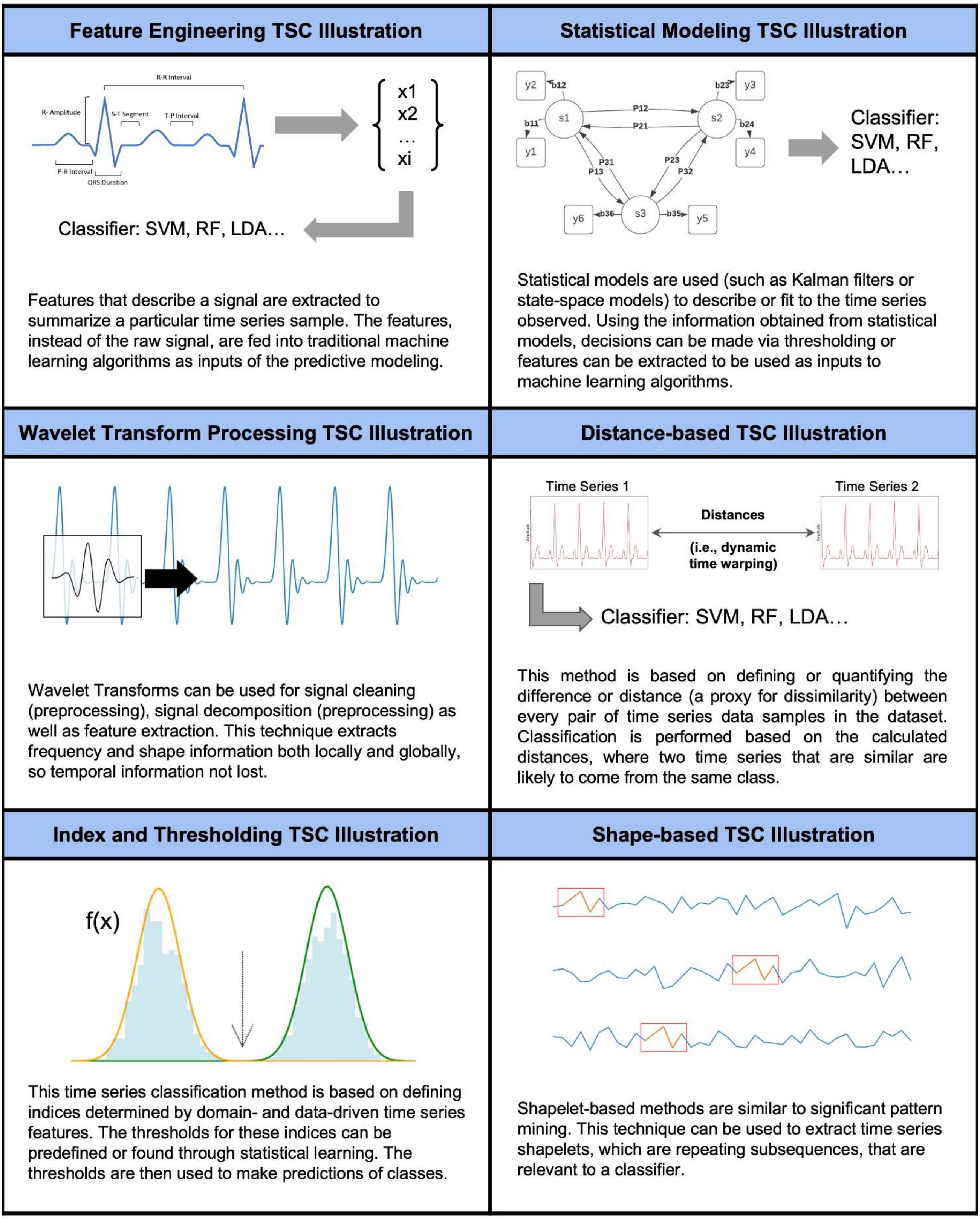
4. Algorithm Summaries

Among the articles reviewed, electroencephalogram (EEG) signals were the most common biosignals investigated. Detailed information about the types of signals investigated in these papers are shown in Figure 2a. Engineered time series features that are fed into widely used machine learning classifier models are the most commonly used technique and most often found achieve the best performance (77 out of 135 articles). Statistical modeling (24 out of 135 articles) algorithms are the second most common. Wavelet-based classification models (8 out of 135 articles) are also common (Figure 2c). Of papers that reported accuracy, 64% achieved accuracy higher than 90%. Of those that reported F1-scores, 70% achieved an F1-score higher than 0.90. Of those that reported AUC-ROC values, 24% achieved AUC-ROC values higher than 0.90. Of those that reported the sensitivity and/or specificity, 54% and 57%, respectively, achieved scores higher than 0.90. Of those that reported Cohen’s Kappa, 43% achieved a Cohen’s Kappa higher than 0.90.



(**a**)

Figure 2. *Cont*.



Feature Engineering TSC Illustration

Features that describe a signal are extracted to summarize a particular time series sample. The features, instead of the raw signal, are fed into traditional machine learning algorithms as inputs of the predictive modeling.

Statistical Modeling TSC Illustration

Statistical models are used (such as Kalman filters or state-space models) to describe or fit to the time series observed. Using the information obtained from statistical models, decisions can be made via thresholding or features can be extracted to be used as inputs to machine learning algorithms.

Wavelet Transform Processing TSC Illustration

Wavelet Transforms can be used for signal cleaning (preprocessing), signal decomposition (preprocessing) as well as feature extraction. This technique extracts frequency and shape information both locally and globally, so temporal information not lost.

Distance-based TSC Illustration

This method is based on defining or quantifying the difference or distance (a proxy for dissimilarity) between every pair of time series data samples in the dataset. Classification is performed based on the calculated distances, where two time series that are similar are likely to come from the same class.

Index and Thresholding TSC Illustration

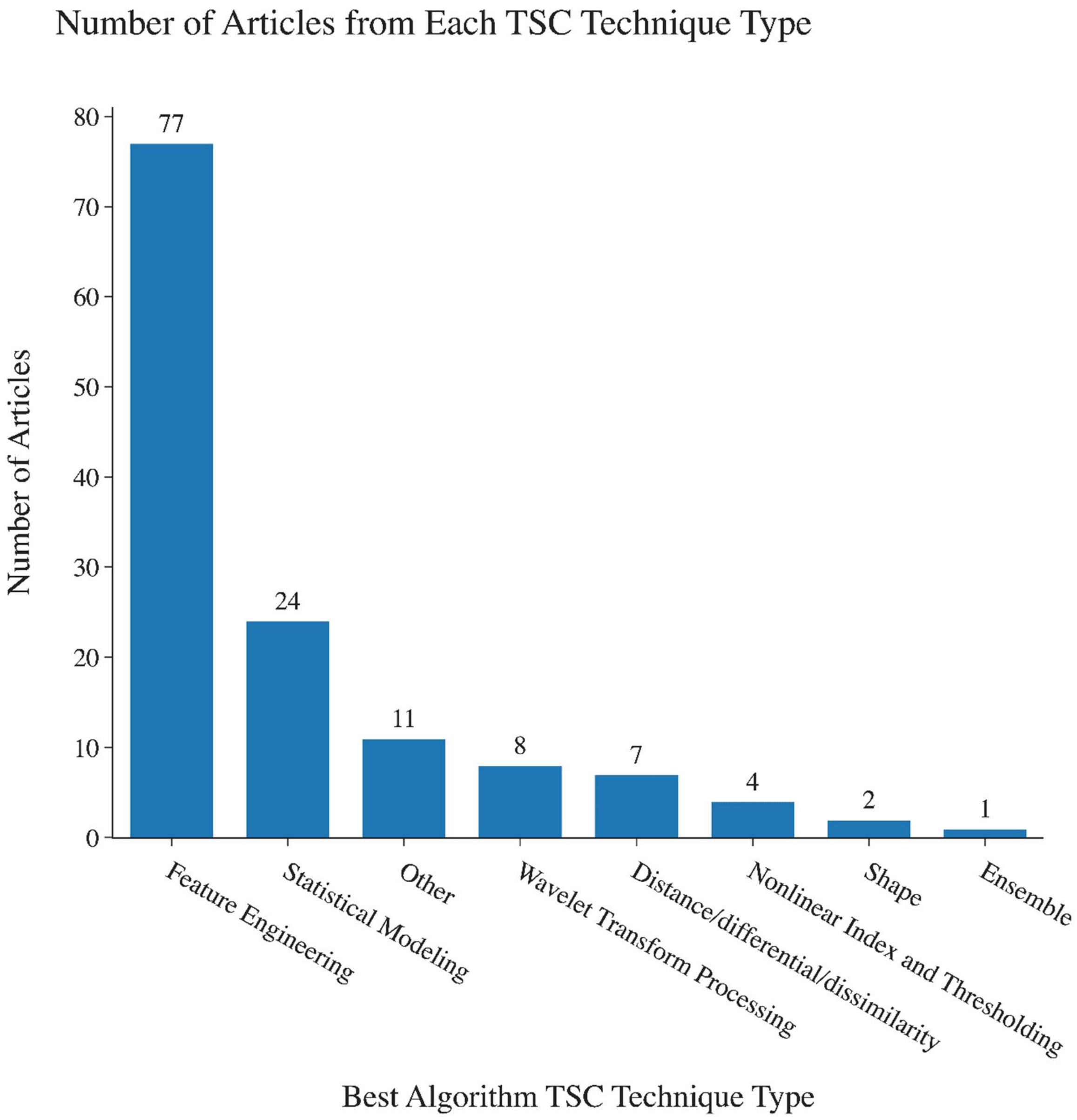
This time series classification method is based on defining indices determined by domain- and data-driven time series features. The thresholds for these indices can be predefined or found through statistical learning. The thresholds are then used to make predictions of classes.

Shape-based TSC Illustration

Shapelet-based methods are similar to significant pattern mining. This technique can be used to extract time series shapelets, which are repeating subsequences, that are relevant to a classifier.

(**b**)

Figure 2. *Cont.*



(**c**)

Figure 2. (**a**). Numbers of papers found in this review focusing on each different biosignal type specified on the horizontal axis. (**b**). Conceptual representation of non-deep learning time series classification modeling types. [1,16–20]. (**c**). Number of articles found for the categories of time series classification methods (horizontal axis) used in biomedical applications.

The classification performance metrics of all the articles were recorded and included in this review, including accuracy, F1-score, Area Under Curve of Receiver Operating Characteristics (AUC-ROC), sensitivity, specificity, and Cohen's Kappa. The accuracy score is the most commonly reported performance metric, with 68% of the articles reporting accuracy scores (Figure 3). All other performance metrics are seldom reported: 30% reported F1-score, 19% reported AUC-ROC, 35% reported specificity, 43% reported sensitivity, and 6% reported Cohen's Kappa. This is concerning because oftentimes using only one or two performance metrics to evaluate a classifier is unreliable and does not tell the whole story of performance [21,22]. In the 135 papers reviewed, 86.7% reported one or more performance metrics, 50.4% reported two or more performance metrics, and 37.8% reported three or more performance metrics. Only 2 papers out of 135 reported more than 6 performance metrics.

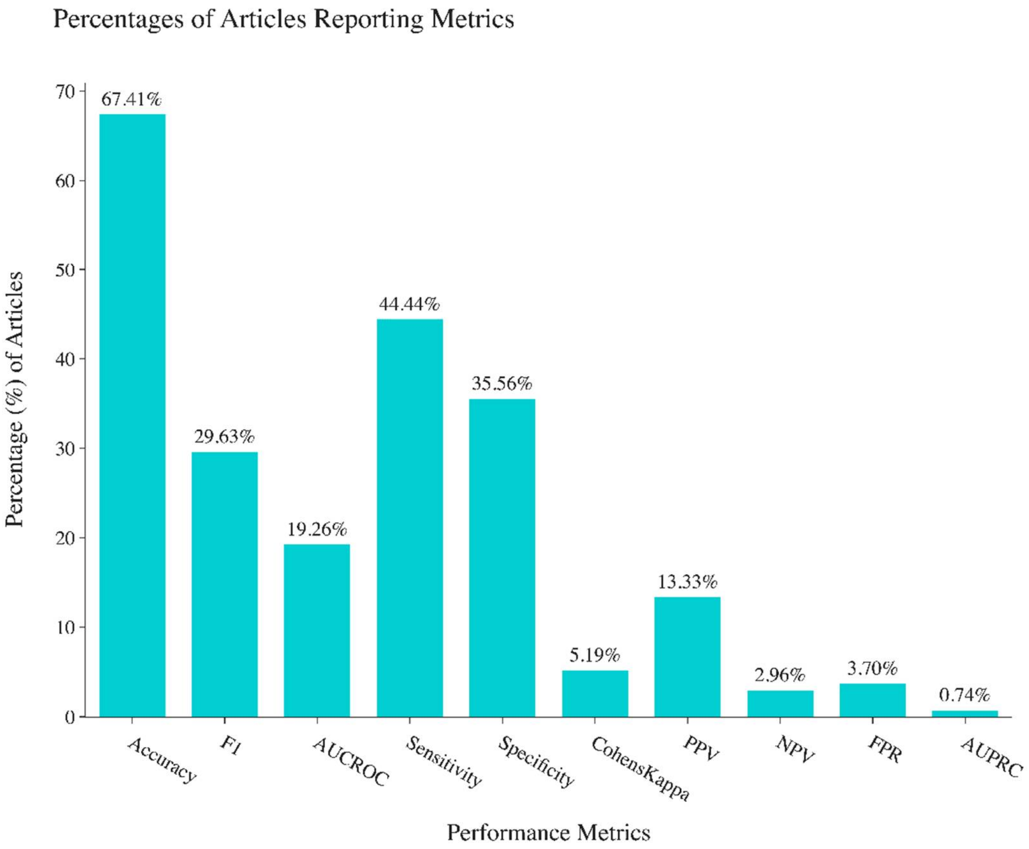


Figure 3. Percentages of performance metrics reported in studies reviewed.

4.1. Preprocessing Methods

In all 135 papers, 50% specifically mentioned the preprocessing methods used. The most common preprocessing method is filtering, which was used mainly for artifact removal or noise reduction. Some other common preprocessing methods include re-sampling (downsampling for lower frequency or upsampling for higher frequency), segmentation, and smoothing. Other common methods are the use of discrete wavelet transform to decompose the original signal into different frequency bands [23–25], the use of continuous wavelet transform to expand the feature space [26], and the use of Fourier transform for signal decomposition and feature extraction [27,28]. There are also intelligent upsampling techniques, such as the use of synthetic data generation for a larger sample during preprocessing [29]. We present a summary of the commonly used preprocessing methods in Supplementary Table S3.

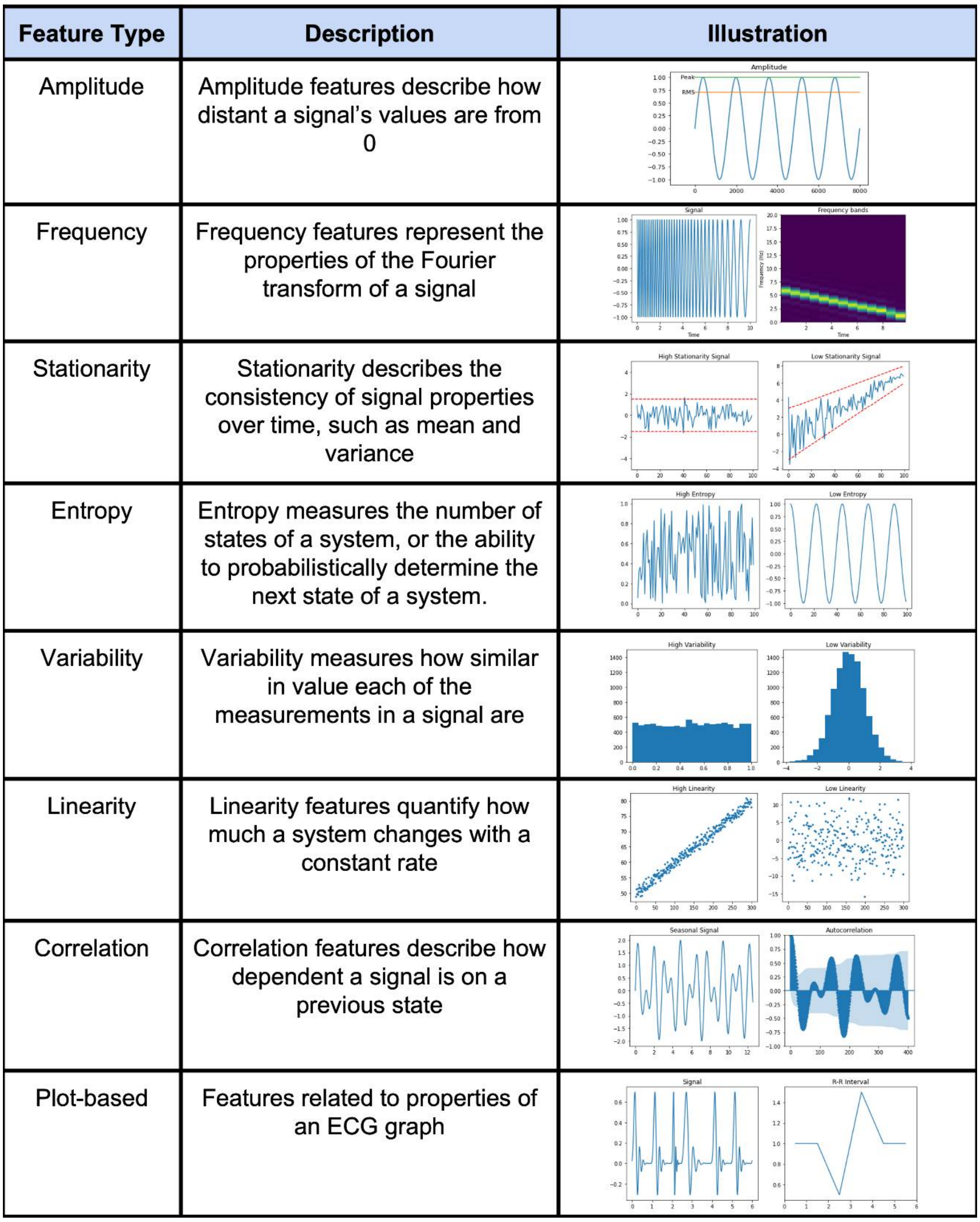
4.2. Feature Engineering Methods

Feature engineering was the most commonly used method of time series classification. The feature engineering pipeline (Figure 1b) usually consists of the following steps:

1. Preprocessing: this step takes raw data as the input and performs some manipulation of the data to return cleaner signals. Common steps include artifact removal, filtering, and segmentation.
2. Signal transformation: this step can be used in preprocessing and also as a precursor to feature extraction. Some manipulation is performed on the signal to represent it in a different space. Common choices are Fourier Transform and wavelet transforms.
3. Feature extraction: in this step, features are extracted from the time series data as a new representation of the original time series.

4. Feature selection: this step selects the features that are the most descriptive, or have the most explanation power. Feature selection is also frequently performed in conjunction with model building.
5. Model selection: the best model is found through hyperparameter tuning and/or comparisons between different types of algorithms.
6. Model validation: performance metrics are calculated for all of the final models. This is frequently done in conjunction with model selection and often using some form of cross-validation.

An example feature engineering technique for a time series is shown in Figure 4. Summary tables of the extracted features for general time series data as well as specific signals (HRV, EEG, etc.) and feature selection methods are presented in Supplementary Table S4a,b. Supplementary Table S5 also presents a summary table for all of the found feature selection methods. In short, feature engineering is used for all signal types across many different applications.



Feature Type	Description	Illustration
Amplitude	Amplitude features describe how distant a signal's values are from 0	Amplitude
Frequency	Frequency features represent the properties of the Fourier transform of a signal	Signal Frequency bands
Stationarity	Stationarity describes the consistency of signal properties over time, such as mean and variance	High Stationarity Signal Low Stationarity Signal
Entropy	Entropy measures the number of states of a system, or the ability to probabilistically determine the next state of a system.	High Entropy Low Entropy
Variability	Variability measures how similar in value each of the measurements in a signal are	High Variability Low Variability
Linearity	Linearity features quantify how much a system changes with a constant rate	High Linearity Low Linearity
Correlation	Correlation features describe how dependent a signal is on a previous state	Seasonal Signal Autocorrelation
Plot-based	Features related to properties of an ECG graph	Signal R-R Interval

Figure 4. Illustration of different types of feature engineering techniques [30].

4.3. Other Methods

Ensemble Methods: Ensemble-based methods are characterized by the connection of multiple algorithmic models that join forces to make the final prediction. These methods may or may not need an additional feature engineering step. Some algorithms that do not necessitate feature engineering in this category are Hierarchical Vote Collective of Transformation-based Ensembles and Bag of Symbolic Fourier Approximation Symbols ensemble algorithms (BOSS) [13]. Newman et al. [31] describe a novel 3-classifier ensemble algorithm for detecting short periods of artificially induced nystagmus (a vision condition in which the eyes make repetitive, uncontrolled movements) from continuous eye movement data. The ensemble of classifiers include a support vector machine (SVM), a linear discriminant analysis (LDA), and boosted trees, and the final classification decision is made by the majority vote. This method reported an accuracy of 0.9877, F1-score of 0.98, sensitivity of 0.9911, and specificity of 0.9863. Elsayed et al. [32] tested eight different state-of-the-art time series classification methods to find the optimal univariant ECG signal classifier. These models are: the Fully Convolutional Network (FCN), Long Short-Term Memory and Fully Convolutional Network (LSTM-FCN) and its attention-based LSTM model (ALSTM-FCN), the Deep Gated Recurrent and Convolutional Network Hybrid Model (GRU-FCN), the Residual Network Mode (ResNet), Multilayered Perceptron model (MLP), Dynamic Time Warping model (DTW), and the noise-reduction-based model, BOSS. The best performance resulted from GRU-FCN, which achieved the highest accuracy in five out of the six datasets that were tested, with a reported accuracy score of 0.92.

State-space Models: State-space models are characterized by the construction of a state and transition model where the transitions are modeled by probabilities. Often, state-space models are most intuitively used for sequence-to-sequence or point-wise classification. For example, She et al. [33] introduced an adaptive transfer learning algorithm to classify and segment events from non-stationary, multi-channel temporal data recorded by an Empatica E4 wristband, including 3-axis accelerometry (ACC), heart rate (HR), skin temperature (TEMP), and electrodermal activity (EDA). Using a multivariate Hidden Markov Model (HMM) and Fisher's Linear Discriminant Analysis (FLDA), the algorithm adaptively adjusts to shifts in the distribution over time, thereby achieving an accuracy of 0.9981 and F1-score of 0.9987. Garcia et al. [34] proposed a method based on dynamic affect (or emotional state) recognition from multimodal physiological signals such as EEG, Electrooculography (EOG), and Electromyography (EMG). This model is based on learning about latent space using Gaussian Process Latent Variable Models (GP-LVM), which maps high-dimensional data (multimodal physiological signals) to a low-dimensional latent space. A support vector classifier is implemented to evaluate the relevance of the latent space features in the affect recognition process, thereby achieving an accuracy of 0.90556.

Shape/Pattern-based: These models are characterized by mining or comparing shapes or patterns in a time or sequence vector. For example, Zhou et al. [35] published an algorithm that can take into consideration the interaction among signals collected at spatiotemporally distinct points, where fuzzy temporal patterns are used to characterize and differentiate between different classes of multichannel EEG data. This algorithm achieved an accuracy of 0.9318 and an F1-score of 0.931, thereby classifying positive vs negative emotion states.

Distance-based: These models calculate the distance (or differences) of time series data vectors. For example, Forestier et al. [36] propose an efficient algorithm to find the optimal partial alignment (optimal subsequence matching) and a prediction system for multivariate signals using maximum a posteriori probability estimation and filtering. This scoring function is based on dynamic time warping. They were able to achieve an accuracy of 0.95, an F1-score of 0.926, and a sensitivity of 0.896.

Other: There are other methodologies that are difficult to characterize. One common method is performed by using statistical modeling of some sort. For example, İşcan et al. [37] published a high performance method to classify and discriminate various ECG patterns (to identify and classify QRS complexes). The model is called LLGMN, which is composed of a Log-Linear Model and a Gaussian Mixture Model (GMM), and gives a posterior probability for the training data. This model was able to achieve the highest accuracy, which was 0.9924.

Another common method is designing a composite metric or index based on domain knowledge or data-driven metrics. For example, Zhou et al. [38] proposed a new algorithm to detect gait events on three walking terrains in real-time based on an analysis of acceleration jerk signals with a time–frequency method to obtain gait parameters, as well as detecting the peaks of jerk signals using peak heuristics. The performance of the newly proposed algorithm was evaluated in eight healthy subjects walking on level ground, upstairs, and downstairs. The mean F1-score was above 0.98 for HS (heel-strike) event detection and 0.95 for TO (toe-off) event detection on the three terrains.

Some articles focus specifically on investigating the wavelet transform and increasing its usefulness for specific use cases. For example, Ji et al. [39] systematically investigated the performances of mother wavelets commonly used in detecting gait events. The overall performance of the Continuous Wavelet Transform (CWT) in detecting the two gait events was significantly different when using various mother wavelets. "Db6" has the highest detection accuracy with the lowest detection time-error, achieving a final accuracy of 1.0. Lu et al. [40] proposed two methods: Discrete Wavelet Transform (DWT) and Extra Trees Classifier, and a personal identification method based on Continuous Wavelet Transform (CWT) and Convolutional Neural Networks (CNN). Nested five-fold cross-validation was used for model selection and model assessment. The CWT method was adopted to uncover feature differences between EMG signals of different subjects. The two methods achieved accuracies of 0.99206 and 0.99203, respectively.

4.4. Interpretation Methods

Model interpretability is a significant aspect of model building. In time series classification for biomedical applications, the interpretation of models that have been built and validated could highlight potential insights into the biomedical phenomenon of interest. Some models have a built-in methodology of interpretation, such as statistical modeling (Hidden Markov Models, Bayesian Models, or ARIMA models) and indices that are informed based on domain knowledge. For many more models with great performance, however, interpretability is a challenge. Only 47 out of the 135 papers reviewed have included some form of interpretation method or model explanation method. Table 2 summarizes the different types of model interpretation methods with descriptions and some examples.

Table 2. Summary of the different types of model interpretation methods discussed or used in each article.

Type of Interpretation Method	Description	Example Papers
Plotting and Annotating Raw Signal	Plotting and annotating raw signals is a widely adopted and useful method for explaining the significance of differentiating features or shapes in time series classification problems. The plots generally consist of a representation of the raw or preprocessed signals in scatter or line plots and highlight the characteristics of the raw or transformed signal, which serves as the differentiating features or shapes for different classes. Some groups have also adopted plotting of preprocessed and transformed signals to present interpretable results. Examples of this method include plotting heart rate values with steps that compare rest and active periods, plotting detected anomalous sequences that are compared against normal sequences, and plotting time series samples in cluster plots after dimension reduction or feature extraction.	[41–46]

Table 2. *Cont.*

Type of Interpretation Method	Description	Example Papers
Visualization of indices over biological/physiological constructs	Instead of plotting against raw signals in 1D, researchers also routinely plot calculated or estimated metrics against 2D or 3D biological constructs, especially when the time series data are signals that represent complicated biological systems, such as brain activities or blood circulation. This interpretation method is very commonly used on electroencephalogram datasets, and examples include a graphical representation of the brain for mean calculated metrics for calm and distressed individuals, as well as a construction of 2D maps of scalp topographies that indicate statistical differences.	[4,26,28,47]
Statistical Analysis/Modeling	Statistical analysis and modeling are used to provide interpretability for not just the models built for classification, but also for clinical application and biomedical understanding. Various plots and tests can be used to demonstrate the relationship between outcomes and certain features or estimated metrics. Example plots are kernel distribution plots, distribution box-plots from statistical models, normality plots, and the visualization of the separability of indices through plotting of the index space. Example analysis tests include variance analysis, normality tests, correlation analysis, and also modeling techniques such as generalized linear models, bivariate random-effects models, and Bayesian hierarchical models.	[26,41,48–53]
Feature weight/importance analysis/ranking	Analysis and visualization of feature importance in a model are very helpful for researchers and clinicians to identify the most useful and important features that contribute to predicting an outcome or influencing a diagnosis. Many time series classification algorithms have built-in methods for feature importance analysis, such as Random Forest, Logistic Regression, and some statistical modeling based classification algorithms. In the pipeline of feature engineering techniques of time series classification, it is often seen that feature selection or dimension reduction are used, and these steps also automatically generate a ranking of feature importance to model building. Additionally, feature importance and ranking can be generated by specific techniques such as Fisher Importance score and Shapley values.	[44,54–60]
Classifier Boundary Plotted against features	Plotting the classifier's boundary in the feature space or lower dimensional space helps to visualize the classifier's ability to differentiate observations from one class to another, i.e., separability. SVM-based classifiers commonly utilize this method for interpretability.	[61]
Index Parameter and Threshold Tuning	Index parameter and threshold tuning is an interpretation method that is usually used in conjunction with classifier building by using a domain-driven approach. Using a domain-driven approach, the researchers typically try to design an index to quantify a biological or physiological phenomenon. The design of the index is usually flexible and can be tuned by changing the parameters used in the index's formula. The threshold of the index is used to differentiate the classes (such as normal vs abnormal conditions, positive vs negative diagnosis). Both the parameters and the threshold of the designed index can be tuned using the existing the dataset, and the classifier's performance metrics can be examined to find the best set of parameters and threshold(s) to achieve the best classifier performances. These parameters and thresholds could also have biomedical significance and meaning relevant for future medical understanding and research. Index analysis can be performed against record length (length of time series), missingness, sample saturation, and time offset.	[51,57,62,63]

Table 2. *Cont.*

Type of Interpretation Method	Description	Example Papers
Channel or Signal Selection	Channel or signal selection is a model building technique but also an interpretation method. Given the prevalence of multivariable time series data in biomedical applications, it is critical for researchers to determine which signals among many, or which channels, can be used for classification. By comparing the classifiers' performances using different signals/channels or specifications of the signals (such as where the sensors are placed), researchers are able to find the best combination that achieves the best performance results, and hence provides interpretability in terms of which signal types or channels are most important for the given biomedical application.	[64–66]
Performance Comparisons Investigating Different Scenarios	Comparing classifier performance metrics when built under different scenarios serves as an interpretation method as well as an experimentation method. Experts in a domain of interest can make sense of why a certain scenario produces the best predictability or algorithm performance, thereby contributing to biomedical understanding and research. For example, accuracy and F1-scores can be compared using datasets collected under different sensor inputs, different user locations, different symbolic or discretization methods, and different data fusion techniques	[60,67]
Bland–Altman plot illustrating the agreement	Bland–Altman plots can be used to evaluate the difference between estimated predictions from the algorithm and the gold standard, thereby providing interpretation of the algorithm's prediction power and potential usefulness as a digital biomarker.	[68]
Deep Learning Network Analysis	Although deep learning models are generally thought of as black box models without easy and direct insight into what the models are doing, there has been recent and impactful research into developing the interpretability of deep learning models and some methods for model explanation. The cited example paper introduces a "global and local explanation". Global explanation means looking at entire classes of data that show which regions of the signal patterns have the most influence for a specific class. Local explanation is the analysis of specific input signals and model outcomes. These methods enable a deeper understanding of the network's behavior, thereby showing the most informative regions that trigger the classification decision and highlighting the possible causes of abnormal physiology or behavior.	[69]

4.5. Best Performing Algorithms

While it is impossible to reasonably declare that one particular type of time series classification algorithm is best for all biomedical applications, it is possible to recommend certain algorithms that achieved great performances and are commonly cited for each different input data type. The algorithm(s) that achieved the best performances are selected and summarized for each of the following most common input signal types: Electrocardiogram (ECG), Electroencephalogram (EEG), Actigraphy, Electromyogram (EMG), Photoplethysmogram (PPG), and Inertial Measurement Unit (IMU). In these papers, wavelet transform processing combined with neural network classifiers achieved the best results for Actigraphy data. (Note: This neural-network-based model is included in the review because it is not exactly deep learning, given our focus on time series specific transformation techniques.) Overall, the statistical modeling classifiers and feature engineering methods performed the best and most consistently for all input signal types. We also observed that wavelet transformation was consistently used as a preprocessing method, feature extraction method, or as an integral part of index development, and furthermore, that it achieved great results.

Electrocardiogram (ECG): Adam et al. [70] introduced a time series classification algorithm that extracts 224 non-linear features and relative wavelet features, selects 15 features ranked by the ReliefF method, and uses the k-nearest neighbor as the classifier for the detection of cardiovascular diseases from electrocardiogram signals. This approach achieved an average accuracy of 0.9927, sensitivity of 0.9974, specificity of 0.9808, and positive predictive value (PPV) of 0.9952. This algorithm can be categorized as feature engineering combined with wavelet transform processing.

While the intention of this review is to find non-deep learning algorithms that can perform very well, some algorithms that cannot be completely classified as just deep learning models are also included, for example, when the neural network architecture is very small with a stronger focus on statistical modeling—or when it's designed with a time series specific transformation. Here, two such algorithms are included for discussion that also perform very well for ECG signals. Iscan et al. [37] present a time series classifier model composed of a Log-Linear model and a Gaussian mixture model—short-handed as LLGMN. This is essentially a neural-network-based model that gives a posteriori probability for the training data. This algorithm was able to achieve a high accuracy of 0.9924. He et al. [71] presented an algorithm using Continuous Wavelet Transform and Convolutional Neural Networks to achieve the best accuracy, which was 0.9923, and an F1-score of 0.994, a sensitivity of 0.9941, and specificity of 0.9891.

Electroencephalogram (EEG): Newman et al. [31] described a novel 3-classifier ensemble algorithm for detecting short periods of artificially induced nystagmus from the long-term eye movement data collected by the CAVA, which achieved an accuracy of 0.9877, F1-score of 0.98, sensitivity of 0.99, and specificity of 0.9963. The frequency domain features were used and calculated using FFT. The ensemble of classifiers include an SVM, an LDA, and boosted trees, and the final classification decision is made by a majority vote. [An efficient automatic arousals detection algorithm in single channel EEG.] Ugur et al. [72] used a simple SVM classifier to detect arousal state. This algorithm was able to achieve an accuracy of 0.982, F1-score of 0.962, sensitivity of 0.9467, and specificity of 0.9933. The features that were extracted were the mean and the variance of the scalogram (CWT squared) coefficients for a range of 16–21 Hz.

Actigraphy: Casado et al. [73] examined various different methods to classify and recognize walking, which was captured by the inertial sensors (accelerometer, gyroscope, and magnetometer) of a mobile phone. The authors examined both feature-based techniques and shape-based techniques. For the shape-based techniques, the authors evaluated the subsequence dynamic time warping, support vectors of an SVM as representative patterns, Partitioning Around Medoids (PAM) as representative patterns, and supervised summarization. The shape-based techniques achieved the best accuracy, which was 0.9535, by using a support vector machine with an rbf kernel. Among the feature-based techniques, the best accuracy was achieved by a Random Forest model with an accuracy score of 0.9531. The Convolutional Neural Network models achieved the best accuracy of 0.9834, even with one input channel (as opposed to nine channels in the other models). Islam et al. [74] published an algorithm that achieved an accuracy of 0.9523, F1-score of 1.0, sensitivity of 0.75, and specificity of 0.8824 using a Random Forest algorithm. The 23 features used are the maximum, minimum, average, standard deviation, variance, coefficients of variations in duration (CVD) of stride, stance and swing intervals, age, approximate entropy (ApEn), weight, height, sex, and gait speed of participants.

Electromyograph (EMG): Lu et al. [40] presented an algorithm using EMG for personal recognition that uses feature engineering and the ExtraTrees Classifer, thereby achieving an accuracy of 0.99206. Meshab et al. [26] presented an algorithm for the prediction of recovery from spinal cord injury using EMG. This algorithm extracts features using the time-domain EMG total power and pattern variability, frequency-domain features computed using Fast Fourier Transform (FFT), Short-Time Fourier Transform (STFT), and Continuous Wavelet Transform (CWT), and makes predictions using kNN, thereby achieving an accuracy of 0.975.

Photoplethysmogram (PPG): She et al. [33] presented an algorithm that uses a multivariate Hidden Markov Model to adaptively learn the data distribution and a Linear Discriminant analysis to classify sleep vs wake from PPG data. This algorithm achieved an average accuracy score of 0.9981 and an F1-score of 0.9987.

Inertial Measurement Unit (IMU): Hemmati et al. [75] presented a wavelet-based algorithm to detect postural transitions. The inertial signal was decomposed using a 4th-order Daubechies Wavelet Transform and the classifier uses subject-specific fixed thresholds (curve length and area under the curve) to achieve an accuracy as high as 0.96. Pham et al. [68] presented an algorithm for the detection of steps using Continuous Wavelet Transform and found the minimum and maximum, thereby achieving an accuracy score of 0.99, sensitivity score of 0.9, specificity score of 0.88, PPV of 0.96, and NPV of 0.73. Martindale et al. [76] presented an algorithm for the prediction of activity levels using a hierarchical Hidden Markov Model, thereby achieving an F1-score of 0.962, a sensitivity score of 0.956, and a specificity score of 0.992.

5. Discussion

While deep learning methods have seen wide usage and high performance in health informatics in recent years, this review demonstrates the utility and power of non-deep learning machine learning algorithms. Many papers were reviewed with a focus on conventional machine learning algorithms that achieved almost perfect performance in classification metrics (i.e., 0.999 in classification accuracy). Compared to deep learning approaches, many conventional machine learning algorithms can be used off-the-shelf, without the researcher needing to rebuild the model architecture and tune a large number of hyperparameters. Conventional machine learning algorithms are also generally easier to train, optimize, and deploy due to their light-weight model (not necessarily needing a large number of parameters as in deep neural networks). This review also serves to identify the non-deep learning time series classification techniques that can serve as a competitive baseline comparator for researchers to understand whether newly designed deep learning networks are truly performing well or not. Among all of the papers reviewed, feature engineering methods followed by off-the-shelf machine learning techniques such as Support Vector Machine and Random Forest are by far the most common. To aid future researchers in building and testing feature engineering algorithms, we have provided an almost exhaustive list of features and transformation techniques that can be applied to longitudinal data in health informatics. A summary of the most common preprocessing methods has been provided, but we do not claim the summary to be exhaustive since preprocessing methods are very frequently domain dependent, and often decisions about preprocessing are made with the researchers' own experiences and discretion with considerations about the different characteristics of each unique dataset. There is, however, a lack of standards in terms of the classification metrics reported, which goes against the best practices of reporting multiple metrics to fully describe the performances of the algorithms tested. A total of 42 out of the 135 papers that were reviewed only reported one metric, and 31 of these 42 papers reported the accuracy score, which is prone to bias [22,77].

5.1. Small Datasets

A pipeline of signal processing, feature engineering, and a classifier of choice were able to achieve high classification performances on datasets that came from small populations (<20). For example, Hong et al. [55] (as mentioned above) published an algorithm to detect drowsiness using EEG, PPG, and ECG signals. The data were collected from 16 healthy subjects, and non-linear features were extracted, selected, and fed into a Random Forest Classifier, thereby achieving an accuracy score of 0.99, F1-score of 0.99, and Cohen's Kappa of 0.985. Among the papers that used small datasets, (discrete or continuous) wavelet transform—as a processing method or feature extraction technique—was very commonly used and very effective, such as for the paper presented by Hemmati et al. [75] (mentioned above), which used data from only 12 subjects. Statistical modeling methods are also very effective as classifiers, such as the paper presented by She et al. [33] (mentioned above), which used data from only 20 subjects. This highlights the benefits and present needs of non-deep learning time series classifiers, especially for biomedical applications where time series data with gold standard labels are difficult to come by.

5.2. Clinical Decision Support

Time series classification models are important for clinical decision support, being supplemented by Electronic Health Records (EHR) or other data that are gathered in the clinical setting to make predictions that could help healthcare providers better dedicate attention and resources, such as mortality rate predictions or early detection of sepsis. Again, a pipeline of preprocessing, feature engineering, and classifier models has been very effecitive, particularly because these kinds of models provide the ease of using domain knowledge in model building and have strong and intuitive interpretability. For example, Nancy et al. [56] presented a Statistical Tolerance Roughset-Induced Decision Tree (STRiD) using features that were extracted for the classification of subjects with hepatitis or thrombosis in a clinical setting—as opposed to without—thereby achieving an accuracy score of 0.915, F1-score of 0.9336, and AUC-ROC score of 0.93.

5.3. Medical Devices

Portable and wearable devices have developed stronger capabilities and gained wider usage over the years. Time series modeling using the continuous data stream that comes from these devices can generate medical insights over long stretches of time and identify digital biomarkers that can serve as a screening tool for common medical conditions [78]. Again, the feature engineering pipeline of time series classification is very commonly used and achieves great results due to the limited computational power and storage space found on these devices. An example of the application of time series classification modeling on medical devices is Newman et al.'s study [31], as presented above, which achieved an accuracy of 0.9877 and F1-score of 0.98.

6. Limitations

While rigorous, our paper selection method would have benefited from a third reviewer to break ties and resolve discrepancies. Furthermore, we were not aware of any existing classification system for categorizing the time series classification algorithms, and thus we developed our own, which may be sub-optimal. It is evident that many papers are difficult to categorize or assign a single category because studies often incorporate multiple different approaches, for example, using Dynamic Time Warping to calculate distances between time series motifs, and subsequently using those distances as input features into a Support Vector Machine. Additionally, although we sought to exclude deep learning approaches through our search term design, some papers examined both deep learning and non-deep learning classification algorithms and we felt compelled to include these papers in our review, both to not exclude the non-deep learning methods, as well as to gain insight into the direct comparison between these two approaches.

7. Conclusions

In conclusion, our group performed this systematic review to survey the landscape of non-deep-learning-based time series classification methods used in biomedical applications. Non-deep learning time series classification techniques can be extremely powerful—given their great algorithm performances—while also allowing for great interpretability. However, this field still lacks standardization for model testing and validation procedures and reporting metrics, which should be addressed to allow for better reproducibility and understanding of the algorithms that are presented by researchers in this field.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/s22208016/s1, Table S1: Search terms for each database; Table S2: Description

of each algorithm type and example papers; Table S3: Summary of time series signal preprocessing methods and example papers; Table S4: (a) Example of feature engineering techniques and papers, (b) Features for common signals and example papers; Table S5: Summary of feature selection methods and example papers. Refs. [79–95] are cited in supplementary materials.

Author Contributions: Writer: W.K.W., I.C., L.H., J.Y., A.S.; Reviewer: W.K.W., G.S., Y.J., A.K., J.Z.S., R.Y.; Advisor: A.R.R., M.M.H.S., J.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We also acknowledge the contribution of Sarah Parks, Duke University Librarian.

Conflicts of Interest: There is no conflict of interest found by the authors.

References

1. Bock, C.; Moor, M.; Jutzeler, C.R.; Borgwardt, K. Machine Learning for Biomedical Time Series Classification: From Shapelets to Deep Learning. In *Artificial Neural Networks*; Cartwright, H., Ed.; Springer: New York, NY, USA, 2021; pp. 33–71. [CrossRef]
2. Fawaz, H.I.; Forestier, G.; Weber, J.; Idoumghar, L.; Muller, P.-A. Deep learning for time series classification: A review. *Data Min. Knowl. Discov.* **2019**, *33*, 917–963. [CrossRef]
3. Li, Y.; Tang, X.; Wang, A.; Tang, H. Probability density distribution of delta RR intervals: A novel method for the detection of atrial fibrillation. *Australas. Phys. Eng. Sci. Med.* **2017**, *40*, 707–716. [CrossRef] [PubMed]
4. García-Martínez, B.; Martínez-Rodrigo, A.; Fernández-Caballero, A.; Moncho-Bogani, J.; Alcaraz, R. Nonlinear predictability analysis of brain dynamics for automatic recognition of negative stress. *Neural Comput. Appl.* **2018**, *32*, 13221–13231. [CrossRef]
5. Tabar, Y.R.; Mikkelsen, K.B.; Rank, M.L.; Hemmsen, M.C.; Kidmose, P. Investigation of low dimensional feature spaces for automatic sleep staging. *Comput. Methods Programs Biomed.* **2021**, *205*, 106091. [CrossRef] [PubMed]
6. Smartwatch penetration 2020. Statista. Available online: https://www.statista.com/statistics/1107874/access-to-smartwatch-in-households-worldwide/ (accessed on 25 July 2022).
7. Xie, J.; Wen, D.; Liang, L.; Jia, Y.; Gao, L.; Lei, J. Evaluating the Validity of Current Mainstream Wearable Devices in Fitness Tracking Under Various Physical Activities: Comparative Study. *JMIR mHealth uHealth* **2018**, *6*, e94. [CrossRef] [PubMed]
8. Guillodo, E.; Lemey, C.; Simonnet, M.; Walter, M.; Baca-García, E.; Masetti, V.; Moga, S.; Larsen, M.; Network, H.; Ropars, J.; et al. Clinical Applications of Mobile Health Wearable–Based Sleep Monitoring: Systematic Review. *JMIR mHealth uHealth* **2020**, *8*, e10733. [CrossRef] [PubMed]
9. Bet, P.; Castro, P.C.; Ponti, M. Fall detection and fall risk assessment in older person using wearable sensors: A systematic review. *Int. J. Med. Informatics* **2019**, *130*, 103946. [CrossRef]
10. Turakhia, M.P.; Desai, M.; Hedlin, H.; Rajmane, A.; Talati, N.; Ferris, T.; Desai, S.; Nag, D.; Patel, M.; Kowey, P.; et al. Rationale and design of a large-scale, app-based study to identify cardiac arrhythmias using a smartwatch: The Apple Heart Study. *Am. Hear. J.* **2018**, *207*, 66–75. [CrossRef] [PubMed]
11. Marcus, G. Deep Learning: A Critical Appraisal. *arXiv* **2018**, arXiv:1801.00631.
12. Fisher, A.J.; Medaglia, J.D.; Jeronimus, B.F. Lack of group-to-individual generalizability is a threat to human subjects research. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E6106–E6115. [CrossRef]
13. Ruiz, A.P.; Flynn, M.; Large, J.; Middlehurst, M.; Bagnall, A. The great multivariate time series classification bake off: A review and experimental evaluation of recent algorithmic advances. *Data Min. Knowl. Discov.* **2020**, *35*, 401–449. [CrossRef] [PubMed]
14. Hossin, M.; Sulaiman, M.N. A Review on Evaluation Metrics for Data Classification Evaluations. *Int. J. Data Min. Knowl. Manag. Process* **2015**, *5*, 1–11. [CrossRef]
15. Davis, J.; Goadrich, M. The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd International Conference on Machine Learning, New York, NY, USA, 25–29 June 2006; pp. 233–240. [CrossRef]
16. The Elements of Statistical Learning. Available online: http://link.springer.com/book/10.1007/978-0-387-84858-7 (accessed on 14 August 2022).
17. Hidden Markov Model. Wikipedia. 18 July 2022. Available online: https://en.wikipedia.org/w/index.php?title=Hidden_Markov_model&oldid=1098931761 (accessed on 14 August 2022).
18. Talebi, S. The Wavelet Transform. Medium. 8 January 2021. Available online: https://towardsdatascience.com/the-wavelet-transform-e9cfa85d7b34 (accessed on 13 February 2022).
19. Regan, M. K Nearest Neighbors & Dynamic Time Warping. 3 August 2022. Available online: https://github.com/markdregan/K-Nearest-Neighbors-with-Dynamic-Time-Warping (accessed on 14 August 2022).

20. Figure 10: Using the Gaussian Mixture Model to Estimate the Threshold. ResearchGate. Available online: https://www.researchgate.net/figure/Using-the-Gaussian-Mixture-Model-to-Estimate-the-Threshold_fig4_307984756 (accessed on 14 August 2022).
21. Sokolova, M.; Japkowicz, N.; Szpakowicz, S. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. In *AI 2006: Advances in Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 1015–1021. [CrossRef]
22. Luque, A.; Carrasco, A.; Martín, A.; de las Heras, A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognit.* **2019**, *91*, 216–231. [CrossRef]
23. Mole, S.S.S.; Sujatha, K. An efficient Gait Dynamics classification method for Neurodegenerative Diseases using Brain signals. *J. Med. Syst.* **2019**, *43*, 245. [CrossRef] [PubMed]
24. Joshi, D.; Khajuria, A.; Joshi, P. An automatic non-invasive method for Parkinson's disease classification. *Comput. Methods Programs Biomed.* **2017**, *145*, 135–145. [CrossRef] [PubMed]
25. Tor, H.T.; Ooi, C.P.; Lim-Ashworth, N.S.; Wei, J.K.E.; Jahmunah, V.; Oh, S.L.; Acharya, U.R.; Fung, D.S.S. Automated detection of conduct disorder and attention deficit hyperactivity disorder using decomposition and nonlinear techniques with EEG signals. *Comput. Methods Programs Biomed.* **2021**, *200*, 105941. [CrossRef]
26. Mesbah, S.; Gonnelli, F.; Angeli, C.A.; El-Baz, A.; Harkema, S.J.; Rejc, E. Neurophysiological markers predicting recovery of standing in humans with chronic motor complete spinal cord injury. *Sci. Rep.* **2019**, *9*, 14474. [CrossRef] [PubMed]
27. Anh, N.X.; Nataraja, R.; Chauhan, S. Towards near real-time assessment of surgical skills: A comparison of feature extraction techniques. *Comput. Methods Progr. Biomed.* **2019**, *187*, 105234. [CrossRef]
28. Durongbhan, P.; Zhao, Y.; Chen, L.; Zis, P.; De Marco, M.; Unwin, Z.C.; Venneri, A.; He, X.; Li, S.; Zhao, Y.; et al. A Dementia Classification Framework Using Frequency and Time-Frequency Features Based on EEG Signals. *IEEE Trans. Neural Syst. Rehabil. Eng. Publ. IEEE Eng. Med. Biol. Soc.* **2019**, *27*, 826–835. [CrossRef]
29. Bhattacharya, S.; Mazumder, O.; Roy, D.; Sinha, A.; Ghose, A. Synthetic Data Generation Through Statistical Explosion: Improving Classification Accuracy of Coronary Artery Disease Using PPG. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 1165–1169. [CrossRef]
30. Walter, S.; Gruss, S.; Limbrecht-Ecklundt, K.; Traue, H.C.; Werner, P.; Al-Hamadi, A.; Diniz, N.; da Silva, G.M.; Andrade, A.O. Automatic pain quantification using autonomic parameters. *Psychol. Neurosci.* **2014**, *7*, 363–380. [CrossRef]
31. Newman, J.L.; Phillips, J.S.; Cox, S.J.; FitzGerald, J.; Bath, A. Automatic nystagmus detection and quantification in long-term continuous eye-movement data. *Comput. Biol. Med.* **2019**, *114*, 103448. [CrossRef] [PubMed]
32. Elsayed, N.; Maida, A.S.; Bayoumi, M. An Analysis of Univariate and Multivariate Electrocardiography Signal Classification. 2019, 396–399. In Proceedings of the 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), Boca Raton, FL, USA, 16–19 December 2019; pp. 396–399. [CrossRef]
33. She, X.; Zhai, Y.; Henao, R.; Woods, C.W.; Chiu, C.; Ginsburg, G.S.; Song, P.X.K.; Hero, A.O. Adaptive Multi-Channel Event Segmentation and Feature Extraction for Monitoring Health Outcomes. *IEEE Trans. Biomed. Eng.* **2020**, *68*, 2377–2388. [CrossRef] [PubMed]
34. Garcia, H.F.; Alvarez, M.A.; Orozco, A.A. Gaussian process dynamical models for multimodal affect recognition. In Proceedings of the 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, 16–20 August 2016; pp. 850–853. [CrossRef]
35. Zhou, P.-Y.; Chan, K.C.C. Fuzzy Feature Extraction for Multichannel EEG Classification. *IEEE Trans. Cogn. Dev. Syst.* **2016**, *10*, 267–279. [CrossRef]
36. Forestier, G.; Petitjean, F.; Riffaud, L.; Jannin, P. Automatic matching of surgeries to predict surgeons' next actions. *Artif. Intell. Med.* **2017**, *81*, 3–11. [CrossRef]
37. Iscan, M.; Yigit, F.; Yilmaz, C. Heartbeat pattern classification algorithm based on Gaussian mixture model. In Proceedings of the 2016 IEEE International Symposium on Medical Measurements and Applications (MeMeA), Benevento, Italy, 15–18 May 2016; pp. 1–6. [CrossRef]
38. Zhou, H.; Ji, N.; Samuel, O.W.; Cao, Y.; Zhao, Z.; Chen, S.; Li, G. Towards Real-Time Detection of Gait Events on Different Terrains Using Time-Frequency Analysis and Peak Heuristics Algorithm. *Sensors* **2016**, *16*, 1634. [CrossRef]
39. Ji, N.; Zhou, H.; Guo, K.; Samuel, O.W.; Huang, Z.; Xu, L.; Li, G. Appropriate Mother Wavelets for Continuous Gait Event Detection Based on Time-Frequency Analysis for Hemiplegic and Healthy Individuals. *Sensors* **2019**, *19*, 3462. [CrossRef]
40. Lu, L.; Mao, J.; Wang, W.; Ding, G.; Zhang, Z. A Study of Personal Recognition Method Based on EMG Signal. *IEEE Trans. Biomed. Circuits Syst.* **2020**, *14*, 681–691. [CrossRef]
41. Liu, J.; Zhao, Y.; Lai, B.; Wang, H.; Tsui, K.L. Wearable Device Heart Rate and Activity Data in an Unsupervised Approach to Personalized Sleep Monitoring: Algorithm Validation. *JMIR mHealth uHealth* **2020**, *8*, e18370. [CrossRef]
42. Cimbalnik, J.; Brinkmann, B.; Kremen, V.; Jurak, P.; Berry, B.; Van Gompel, J.; Stead, M.; Worrell, G. Physiological and pathological high frequency oscillations in focal epilepsy. *Ann. Clin. Transl. Neurol.* **2018**, *5*, 1062–1076. [CrossRef]
43. Ren, H.; Ye, Z.; Li, Z. Anomaly detection based on a dynamic Markov model. *Inf. Sci.* **2017**, *411*, 52–65. [CrossRef]
44. Elden, R.H.; Ghoneim, V.F.; Al-Atabany, W. A computer aided diagnosis system for the early detection of neurodegenerative diseases using linear and non-linear analysis. In Proceedings of the 2018 IEEE 4th Middle East Conference on Biomedical Engineering (MECBME), Tunis, Tunisia, 28–30 March 2018; pp. 116–121. [CrossRef]

45. Heartbeat Classification Using Abstract Features From the Abductive Interpretation of the ECG | IEEE Journals & Magazine | IEEE Xplore. Available online: https://ieeexplore-ieee-org.proxy.lib.duke.edu/document/7750556 (accessed on 6 February 2022).
46. Gupta, R.; Kundu, P. Dissimilarity factor based classification of inferior myocardial infarction ECG. In Proceedings of the 2016 IEEE First International Conference on Control, Measurement and Instrumentation (CMI), Kolkata, India, 8–10 January 2016; pp. 229–233. [CrossRef]
47. Mohammadi-Ghazi, R.; Marzouk, Y.M.; Büyüköztürk, O. Conditional classifiers and boosted conditional Gaussian mixture model for novelty detection. *Pattern Recognit.* **2018**, *81*, 601–614. [CrossRef]
48. Reamaroon, N.; Sjoding, M.W.; Lin, K.; Iwashyna, T.J.; Najarian, K. Accounting for Label Uncertainty in Machine Learning for Detection of Acute Respiratory Distress Syndrome. *IEEE J. Biomed. Heal. Informatics* **2018**, *23*, 407–415. [CrossRef] [PubMed]
49. Park, G.H.; Kim, S.J.; Cho, Y.S. Development of a voiding diary using urination recognition technology in mobile environment. *J. Exerc. Rehabil.* **2020**, *16*, 529–533. [CrossRef] [PubMed]
50. David, S.; Machado, J.; Inácio, C.; Valentim, C. A combined measure to differentiate EEG signals using fractal dimension and MFDFA-Hurst. *Commun. Nonlinear Sci. Numer. Simul.* **2020**, *84*, 105170. [CrossRef]
51. Li, M.; Tian, S.; Sun, L.; Chen, X. Gait Analysis for Post-Stroke Hemiparetic Patient by Multi-Features Fusion Method. *Sensors* **2019**, *19*, 1737. [CrossRef]
52. Gunnarsdottir, K.; Sadashivaiah, V.; Kerr, M.; Santaniello, S.; Sarma, S.V. Using demographic and time series physiological features to classify sepsis in the intensive care unit. In Proceedings of the 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, 16–20 August 2016; pp. 778–782. [CrossRef]
53. Mertzanis, L.; Panotonoulou, A.; Skoularidou, M.; Kontoyiannis, I. Deep Tree Models for 'Big' Biological Data. In Proceedings of the 2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Kalamata, Greece, 25–28 June 2018; pp. 1–5. [CrossRef]
54. El-Rashidy, N.; El-Sappagh, S.; Abuhmed, T.; Abdelrazek, S.M.; El-Bakry, H.M. Intensive Care Unit Mortality Prediction: An Improved Patient-Specific Stacking Ensemble Model. *IEEE Access* **2020**, *8*, 133541–133564. [CrossRef]
55. Hong, S.; Kwon, H.; Choi, S.H.; Park, K.S. Intelligent system for drowsiness recognition based on ear canal electroencephalography with photoplethysmography and electrocardiography. *Inf. Sci.* **2018**, *453*, 302–322. [CrossRef]
56. Nancy, J.Y.; Khanna, N.H.; Kannan, A. A bio-statistical mining approach for classifying multivariate clinical time series data observed at irregular intervals. *Expert Syst. Appl.* **2017**, *78*, 283–300. [CrossRef]
57. Miao, B.; Guan, J.; Zhang, L.; Meng, Q.; Zhang, Y. Automated Epileptic Seizure Detection Method Based on the Multi-attribute EEG Feature Pool and mRMR Feature Selection Method. In Proceedings of the International Conference on Computational Science—ICCS 2019, Faro, Portugal, 12–14 June 2019; Rodrigues, J.M.F., Cardoso, P.J.S., Monteiro, J., Lam, R., Krzhizhanovskaya, V.V., Lees, M.H., Dongarra, J.J., Sloot, P.M.A., Eds.; Springer: Cham, Switzerland, 2019; Volume 11538, pp. 45–59. [CrossRef]
58. Orphanou, K.; Dagliati, A.; Sacchi, L.; Stassopoulou, A.; Keravnou, E.; Bellazzi, R. Incorporating repeating temporal association rules in Naïve Bayes classifiers for coronary heart disease diagnosis. *J. Biomed. Inform.* **2018**, *81*, 74–82. [CrossRef]
59. Lacson, R.C.; Baker, B.; Suresh, H.; Andriole, K.; Szolovits, P.; Lacson, E. Use of machine-learning algorithms to determine features of systolic blood pressure variability that predict poor outcomes in hypertensive patients. *Clin. Kidney J.* **2018**, *12*, 206–212. [CrossRef]
60. Ozdenizci, O.; Cumpanasoiu, C.; Mazefsky, C.; Siegel, M.; Erdoğgmus, D.; Ioannidis, S.; Goodwin, M.S. Time-Series Prediction of Proximal Aggression Onset in Minimally-Verbal Youth with Autism Spectrum Disorder Using Physiological Biosignals. In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 18–21 July 2018; pp. 5745–5748. [CrossRef]
61. Lotfan, S.; Shahyad, S.; Khosrowabadi, R.; Mohammadi, A.; Hatef, B. Support vector machine classification of brain states exposed to social stress test using EEG-based brain network measures. *Biocybern. Biomed. Eng.* **2018**, *39*, 199–213. [CrossRef]
62. Ródenas, J.; García, M.; Alcaraz, R.; Rieta, J.J. Combined Nonlinear Analysis of Atrial and Ventricular Series for Automated Screening of Atrial Fibrillation. *Complexity* **2017**, *2017*, 2163610. [CrossRef]
63. Cuesta-Frau, D.; Novák, D.; Burda, V.; Molina-Picó, A.; Vargas, B.; Mraz, M.; Kavalkova, P.; Benes, M.; Haluzik, M. Characterization of Artifact Influence on the Classification of Glucose Time Series Using Sample Entropy Statistics. *Entropy* **2018**, *20*, 871. [CrossRef] [PubMed]
64. Zdravevski, E.; Lameski, P.; Trajkovik, V.; Kulakov, A.; Chorbev, I.; Goleva, R.; Pombo, N.; Garcia, N. Improving Activity Recognition Accuracy in Ambient-Assisted Living Systems by Automated Feature Engineering. *IEEE Access* **2017**, *5*, 5262–5280. [CrossRef]
65. Peng, P.; Wei, H.; Xie, L.; Song, Y. Epileptic Seizure Prediction in Scalp EEG Using an Improved HIVE-COTE Model. In Proceedings of the 2020 39th Chinese Control Conference (CCC), Shenyang, China, 27–29 July 2020; pp. 6450–6457. [CrossRef]
66. Li, X.; Zhang, Y.; Jiang, F.; Zhao, H. A novel machine learning unsupervised algorithm for sleep/wake identification using actigraphy. *Chronobiol. Int.* **2020**, *37*, 1002–1015. [CrossRef] [PubMed]
67. Bragança, H.; Colonna, J.G.; Lima, W.S.; Souto, E. A Smartphone Lightweight Method for Human Activity Recognition Based on Information Theory. *Sensors* **2020**, *20*, 1856. [CrossRef]
68. Pham, M.H.; Elshehabi, M.; Haertner, L.; Del Din, S.; Srulijes, K.; Heger, T.; Synofzik, M.; Hobert, M.A.; Faber, G.S.; Hansen, C.; et al. Validation of a Step Detection Algorithm during Straight Walking and Turning in Patients with Parkinson's Disease and Older Adults Using an Inertial Measurement Unit at the Lower Back. *Front. Neurol.* **2017**, *8*, 457. [CrossRef]

69. Ivaturi, P.; Gadaleta, M.; Pandey, A.C.; Pazzani, M.; Steinhubl, S.R.; Quer, G. A Comprehensive Explanation Framework for Biomedical Time Series Classification. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 2398–2408. [CrossRef]
70. Adam, M.; Oh, S.L.; Sudarshan, V.K.; Koh, J.E.; Hagiwara, Y.; Tan, J.H.; Tan, R.S.; Acharya, U.R. Automated characterization of cardiovascular diseases using relative wavelet nonlinear features extracted from ECG signals. *Comput. Methods Programs Biomed.* **2018**, *161*, 133–143. [CrossRef]
71. He, R.; Wang, K.; Zhao, N.; Liu, Y.; Yuan, Y.; Li, Q.; Zhang, H. Automatic Detection of Atrial Fibrillation Based on Continuous Wavelet Transform and 2D Convolutional Neural Networks. *Front. Physiol.* **2018**, *9*, 1206. [CrossRef]
72. Ugur, T.K.; Erdamar, A. An efficient automatic arousals detection algorithm in single channel EEG. *Comput. Methods Programs Biomed.* **2019**, *173*, 131–138. [CrossRef] [PubMed]
73. Casado, F.E.; Rodríguez, G.; Iglesias, R.; Regueiro, C.V.; Barro, S.; Canedo-Rodríguez, A. Walking Recognition in Mobile Devices. *Sensors* **2020**, *20*, 1189. [CrossRef] [PubMed]
74. Islam, R.; Pavel, S.R.; Tunaz, S.A. Neurodegenerative Disease Classification Using Gait Signal Features and Random Forest Classifier. In Proceedings of the 2019 4th International Conference on Electrical Information and Communication Technology, (EICT), Khulna, Bangladesh, 20–22 December 2019; pp. 1–4. [CrossRef]
75. Hemmati, S.; Wade, E. Detecting postural transitions: A robust wavelet-based approach. In Proceedings of the 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, 16–20 August 2016; pp. 3704–3707. [CrossRef]
76. Martindale, C.F.; Hoenig, F.; Strohrmann, C.; Eskofier, B.M. Smart Annotation of Cyclic Data Using Hierarchical Hidden Markov Models. *Sensors* **2017**, *17*, 2328. [CrossRef] [PubMed]
77. Canbek, G.; Temizel, T.T.; Sagiroglu, S. BenchMetrics: A systematic benchmarking method for binary classification performance metrics. *Neural Comput. Appl.* **2021**, *33*, 14623–14650. [CrossRef]
78. Bent, B.; Wang, K.; Grzesiak, E.; Jiang, C.; Qi, Y.; Jiang, Y.; Cho, P.; Zingler, K.; Ogbeide, F.I.; Zhao, A.; et al. The digital biomarker discovery pipeline: An open-source software platform for the development of digital biomarkers using mHealth and wearables data. *J. Clin. Transl. Sci.* **2020**, *5*. [CrossRef] [PubMed]
79. Shi, Y.; Li, F.; Liu, T.; Beyette, F.R.; Song, W. Dynamic Time-frequency Feature Extraction for Brain Activity Recognition. In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 18–21 July 2018; pp. 3104–3107. [CrossRef]
80. Zarei, A.; Asl, B.M. Automatic Detection of Obstructive Sleep Apnea Using Wavelet Transform and Entropy-Based Features From Single-Lead ECG Signal. *IEEE J. Biomed. Health Inform.* **2018**, *23*, 1011–1021. [CrossRef]
81. Liu, L.; Wang, H.; Li, H.; Liu, J.; Qiu, S.; Zhao, H.; Guo, X. Ambulatory Human Gait Phase Detection Using Wearable Inertial Sensors and Hidden Markov Model. *Sensors* **2021**, *21*, 1347. [CrossRef]
82. Zhang, C.; Chen, Y.; Yin, A.; Wang, X. Anomaly detection in ECG based on trend symbolic aggregate approximation. *Math. Biosci. Eng.* **2019**, *16*, 2154–2167. [CrossRef]
83. Lee, S.X.; Leemaqz, S.Y. Automated Wrist Pulse diagnosis of Pancreatitis via Autoregressive Discriminant Models. 2017, pp. 1262–1266. Available online: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85080915306&partnerID=40&md5=f92bd168c60aae1a6ffbda888f5663f0 (accessed on 15 August 2022).
84. Bagattini, F.; Karlsson, I.; Rebane, J.; Papapetrou, P. A classification framework for exploiting sparse multi-variate temporal features with application to adverse drug event detection in medical records. *BMC Med Informatics Decis. Mak.* **2019**, *19*, 7. [CrossRef]
85. Campbell, E.; Phinyomark, A.; Scheme, E. Feature Extraction and Selection for Pain Recognition Using Peripheral Physiological Signals. *Front. Neurosci.* **2019**, *13*, 437. [CrossRef]
86. Jovic, A.; Brkic, K.; Krstacic, G. Detection of congestive heart failure from short-term heart rate variability segments using hybrid feature selection approach. *Biomed. Signal Process. Control* **2019**, *53*, 101583. [CrossRef]
87. Nawaz, R.; Cheah, K.H.; Nisar, H.; Yap, V.V. Comparison of different feature extraction methods for EEG-based emotion recognition. *Biocybern. Biomed. Eng.* **2020**, *40*, 910–926. [CrossRef]
88. Haddi, Z.; Ananou, B.; Trardi, Y.; Ouladsine, M.; Pons, J.-F.; Delliaux, S.; Deharo, J.-C. Relevance Vector Machine as Data-Driven Method for Medical Decision Making. In Proceedings of the 2019 18th European Control Conference (ECC), Naples, Italy, 25–28 June 2019; pp. 1011–1016. [CrossRef]
89. Wickramasuriya, D.S.; Tessmer, M.K.; Faghih, R.T. Facial Expression-Based Emotion Classification using Electrocardiogram and Respiration Signals. In Proceedings of the 2019 IEEE Healthcare Innovations and Point of Care Technologies (HI-POCT), Bethesda, MD, USA, 20–22 November 2019; pp. 9–12. [CrossRef]
90. Malik, A.R.; Boger, J. Zero-Effort Ambient Heart Rate Monitoring Using Ballistocardiography Detected Through a Seat Cushion: Prototype Development and Preliminary Study. *JMIR Rehabilitation Assist. Technol.* **2021**, *8*, e25996. [CrossRef] [PubMed]
91. Liu, N.; Sun, M.; Wang, L.; Zhou, W.; Dang, H.; Zhou, X. A support vector machine approach for AF classification from a short single-lead ECG recording. *Physiol. Meas.* **2018**, *39*, 064004. [CrossRef] [PubMed]
92. Goshvarpour, A. Evaluation of Novel Entropy-Based Complex Wavelet Sub-bands Measures of PPG in an Emotion Recognition System. *J. Med. Biol. Eng.* **2020**, *40*, 451–461. [CrossRef]
93. Kolodziej, M.; Majkowski, A.; Rak, R.J.; Rysz, A.; Marchel, A. Decision Support System For Epileptogenic Zone Location during Brain Resection. *Metrol. Meas. Syst.* **2018**, *25*, 15–32. [CrossRef]

94. Hu, Y.; An, W.; Subramanian, R.; Zhao, N.; Gu, Y.; Wu, W. *Faster Clinical Time Series Classification with Filter Based Feature Engineering Tree Boosting Methods*; Springer: Cham, Switzerland, 2021; Volume 914, p. 260. [CrossRef]
95. Mumtaz, W.; Xia, L.; Yasin, M.A.M.; Ali, S.S.A.; Malik, A.S. A wavelet-based technique to predict treatment outcome for Major Depressive Disorder. *PLoS ONE* **2017**, *12*, e0171409. [CrossRef]