axioms

MDPI

Article

An MDL-Based Wavelet Scattering Features Selection for Signal Classification

Vittoria Bruni *, Maria Lucia Cardinali and Domenico Vitulano

Department of Basic and Applied Sciences for Engineering, Sapienza Rome University, Via Antonio Scarpa 16, 00161 Rome, Italy; marialucia.cardinali@uniroma1.it (M.L.C.); domenico.vitulano@uniroma1.it (D.V.)
* Correspondence: vittoria.bruni@uniroma1.it

Abstract: Wavelet scattering is a redundant time-frequency transform that was shown to be a powerful tool in signal classification. It shares the convolutional architecture with convolutional neural networks, but it offers some advantages, including faster training and small training sets. However, it introduces some redundancy along the frequency axis, especially for filters that have a high degree of overlap. This naturally leads to a need for dimensionality reduction to further increase its efficiency as a machine learning tool. In this paper, the Minimum Description Length is used to define an automatic procedure for optimizing the selection of the scattering features, even in the frequency domain. The proposed study is limited to the class of uniform sampling models. Experimental results show that the proposed method is able to automatically select the optimal sampling step that guarantees the highest classification accuracy for fixed transform parameters, when applied to audio/sound signals.

Keywords: signal classification; minimum description length; support vector machine; wavelet scattering

Citation: Bruni, V.; Cardinali, M.L.; Vitulano, D. An MDL-Based Wavelet Scattering Features Selection for Signal Classification. *Axioms* **2022**, *11*, 376. https://doi.org/10.3390/axioms11080376

Academic Editor: Hans J. Haubold

Received: 17 June 2022
Accepted: 28 July 2022
Published: 30 July 2022

1. Introduction

Wavelet scattering [1–3] is a time-frequency transform that is able to better represent signal characteristics due to the use of a recursive chain. The latter consists of a constant-Q factor wavelet decomposition, a non-linear operation (namely absolute value) and a lowpass averaging filtering for each layer. It is a deep convolutional operator where filters are given instead of being learnt. The Wavelet Scattering Transform (WST) was originally derived from the MEL spectrum decomposition for audio/speech signals processing. It is shift invariant, stable to deformations and non-expansive; as a result, the depth of the network can be limited, as most of the signal energy is concentrated in the first layers. In addition, it allows for a fast implementation. Even though each task requires ad hoc neural network architectures, WST provides useful features that can be an optimal input for specific classifiers or for Convolutional Neural Networks (CNN) themselves [4–9], especially for sound signals. In fact, it overcomes some limitations of Mel Frequency Cepstral coefficients (MFCC) thanks to the CNN-like structure; on the other hand, it allows us to reduce the depth of a deep neural network (DNN) thanks to the compact representation of the significant signal time-frequency structures. For example, for acoustic scenes classification, WST can work better than the baseline CNN when properly combined with a specific classifier—Support Vector Machine (SVM) is used in [4], while two ensemble classifiers are employed in [6]. Similar conclusions are drawn in [10], where WST and SVM are used to successfully classify alcoholic EEG signals, resulting a compelling alternative to CNN-based classification. On the contrary, hybrid architectures, i.e., WST as input for a CNN, guarantee a significant reduction in the number of parameters to be learnt, as shown in [5], where this hybrid architecture has been successfully exploited for speaker identification using a small number of samples as training set.

In CNN architectures, stride is one of the parameters to be set. It is necessary to reduce the data to process at each layer, reducing the computational complexity and eliminating some redundancies that can make the training process more complicated and misleading. While stride is automatically applied by WST in the time domain, the intrinsic redundancy of the transform in the frequency domain could provide too much information, which can be discarded in some cases without affecting the final result. With regard to this point, some papers studied the influence of each layer of the transform in the classification process. In particular, in the pioneering and seminal papers [1,11], the dependence of the classification error on the number of layers has been analysed, and it has been shown that the error does not decrease significantly when using a number of layers greater than three. The more recent study presented in [12] gave evidence of the benefit of using normalized scattering coefficients by exploiting their natural parent–child relationships. Based on the standard data reduction problem [13–16], some others approaches tried to preserve useful scattering coefficients, such as, for example, [17–19]. In this case, Principal Component Analysis (PCA), multidimensional scaling (MDS) and random sampling have been used to reduce the dimension of the scattering feature matrix, while guaranteeing nearly comparable classification accuracy. More precisely, in [18], the problem of arrhythmia classification in ECG signals has been addressed; PCA has been combined with some classifiers, including neural network, probabilistic neural network, and the k-nearest neighbour (kNN), and it has been shown that the last one achieves the best performance. In [17] a twin support vector machine (TWSVM) has been used to classify ECG signals from the wavelet scattering feature matrix, whose dimension has been reduced using MDS. MDS provided more significant features than PCA, while TWSWM contributed to speed up the classification step. Finally, in [19] a random selection of scattering coefficients has been used to reducing 1/4 of the dimension of the feature matrix. Despite the high classification rates, the aforementioned methods require some parameters to be predefined, such as the number of features to preserve, the sampling step or the number of layers. As a consequence, specific criteria for feature selection are required to fully exploit the advantages of the proposed approaches. Feature selection is a widely investigated topic; see [13,20] for a complete review. Briefly speaking, it consists of selecting a subset of features which can efficiently describe the input data while neglecting irrelevant or redundant information but still providing good predictions (such as, for example good classification rates). Feature selection methods can be split into three main classes: filter methods, wrapper methods and embedded methods. The former exploit a specific criterion for ranking the features, from the most to the least significant, and consist of preprocessing of the classification/prediction step. On the contrary, wrapper methods use the performance of the predictor as feature selection criterion. Finally, embedded methods try to combine the advantages of the two aforementioned classes. Independent of the class, the desired goal for a feature selection method is to select those significant and not redundant features with the least computational burden. That is why filter methods are the most popular and widely investigated [20].

Based on the considerations above, this paper investigates a preprocessing method for wavelet scattering coefficients that are able to optimize the learning process in terms of time and/or accuracy. It consists of a uniform sampling along the frequency axis to be applied just before running the classifier. An automatic procedure for the estimation of the best sampling step is proposed. It estimates the uniform sampling of the feature matrix that is able to provide the best classification results for fixed transform settings (Q factors and number of layers). The Minimum Description Length (MDL) [21,22] is used for the automatic best model selection by looking at the compression cost of the analysed sequences. SVM [23] is then used for classification on the basis of the selected model.

Experimental results show that the advantageousness of the proposed approach is twofold. It defines a preprocessing method that is able to optimize the learning process in terms of computing time and/or accuracy, and it introduces the first study concerning an

optimization procedure that depends on the entropy of the layers and that may be directly included in NN architectures in the future.

The remainder of the paper is as follows. The next section provides a brief introduction to the wavelet scattering transform and the minimum description length; then, it describes how they have been combined in the proposed feature-selection-based method. Section 3 presents some experimental results concerning classification of signals through SVM based procedures. Finally, Section 4 draws some conclusions.

2. The Proposed Method

This section introduces the adopted notation by giving a brief description of WST and MDL; then, it presents the details of the proposed method.

2.1. Wavelet Scattering

Wavelet scattering is a non-linear multiscale transform that has a tree structure, such as the one in Figure 1. It consists of a recursive application of proper band-pass filters, but each convolution is followed by a non-linear operation: the absolute value. Each level of the tree consists of the application of a classical redundant filter bank with a predefined Q factor. The scattering coefficients are obtained by lowpass filtering the absolute value of the output of the filter bank, and they are the ones that are retained by the transform. More precisely, the zeroth-order scattering coefficients (layer 0) are defined as

$$S_0(t) = f * \phi(t), \tag{1}$$

where f denotes the analysed signal that depends on the time variable t, ϕ is a lowpass filter, while $*$ denotes the convolution product. The ***zeroth-order layer*** is therefore the row vector $\mathbf{S}_0$, which is composed of N_t temporal samples, as defined in Equation (1).

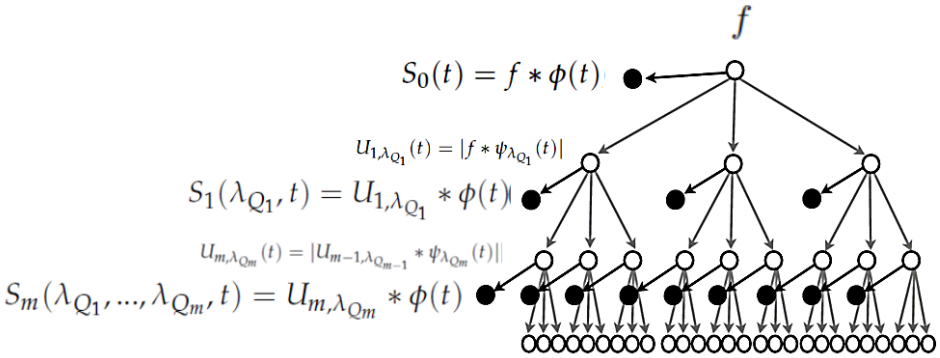

Figure 1. Wavelet Scattering decomposition tree.

The first-order coefficients (first layer) still consist of a lowpass filtering operation that is applied to the absolute value of the output of a Q_1-factor high-pass wavelet filter bank. More precisely, by denoting with $\psi_{\lambda_{Q_1}}$ the temporal wavelet filter dilated by λ_{Q_1}, and with Λ_{Q_1} the set of scaling coefficients that are defined according to the octave resolution Q_1, we have

$$S_1(\lambda_{Q_1}, t) = U_{1,\lambda_{Q_1}} * \phi(t), \qquad \lambda_{Q_1} \in \Lambda_{Q_1} \tag{2}$$

with

$$U_{1,\lambda_{Q_1}}(t) = |f * \psi_{\lambda_{Q_1}}(t)|, \tag{3}$$

where $|\cdot|$ denotes the absolute value. Let $N_{Q_1} = \#\Lambda_1$ be the cardinality of the set Λ_{Q_1}; that is, the number of filters used in the filter bank, then the ***first-order layer*** $\mathbf{S}_1$ is the matrix whose dimension is $N_{Q_1} \times N_t$ and whose rows are as defined in Equation (2).

Accordingly, the m-th layer coefficients are

$$S_m(\lambda_{Q_1}, \ldots, \lambda_{Q_m}, t) = U_{m,\lambda_{Q_m}} * \phi(t), \qquad \lambda_{Q_m} \in \Lambda_{Q_m}, \tag{4}$$

with

$$U_{m,\lambda_{Q_m}}(t) = |U_{m-1,\lambda_{Q_{m-1}}} * \psi_{\lambda_{Q_m}}(t)|. \tag{5}$$

The ***mth-order layer*** $\mathbf{S}_m$ is the matrix, whose rows are defined as in Equation (4), which has dimension $N_{Q_m} \times N_t$, where $N_{Q_m} = \#\Lambda_m$ depends on the number of filters required by the Q_m-filter bank and their overlap with the Q_{m-1}-filter bank.

WST of f is, therefore, the collection of the layers $\mathbf{S}_0, \mathbf{S}_1, ..., \mathbf{S}_m$. More precisely, it is a $N \times N_t$ matrix, with

$$N = 1 + \sum_{k=1}^{m} N_{Q_k}, \tag{6}$$

and consists of the columnwise aggregation of the matrices $\mathbf{S}_0, \mathbf{S}_1, ..., \mathbf{S}_m$; that is

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_0 \\ \mathbf{S}_1 \\ \vdots \\ \mathbf{S}_m \end{bmatrix}. \tag{7}$$

WST is highly redundant, and its redundancy depends on the sequence of Q factors. The latter is a critical issue, as it strictly depends on the analysed signal; in particular, it causes a faster or slower energy decrease as the number of layers increases [1]. However, the selection of the best sequence of Q factors is out of the scope of this paper. On the contrary, for a fixed number of layers, we are interested in reducing the number of scattering coefficients, as they refer to overlapping frequency bands. The rule proposed in this paper is the uniform sampling along the frequency axis. The latter acts as a post-processing operation, and it is applied to the whole WST.

To better decorrelate scattering coefficients, parent–child normalization can be applied [1,3], and the logarithm of the corresponding value can be retained, i.e., $\forall\ t$ and $\forall\ \lambda_{Q_k}$

$$\begin{cases} \tilde{S}_0(t) = log(S_0(t)) \\ \tilde{S}_k(\lambda_{Q_1}, ..., \lambda_{Q_k}, t) = log \frac{S_k(\lambda_{Q_1},...,\lambda_{Q_k},t)}{S_{k-1}(\lambda_{Q_1},...,\lambda_{Q_{k-1}},t)}, \quad k = 1, ..., m \end{cases} \tag{8}$$

As a result, the normalized scattering transform is the $N \times N_t$ matrix

$$\tilde{\mathbf{S}} = \begin{bmatrix} \tilde{\mathbf{S}}_0 \\ \tilde{\mathbf{S}}_1 \\ \vdots \\ \tilde{\mathbf{S}}_m \end{bmatrix}. \tag{9}$$

The latter usually guarantees better classification results [12,18,24].

2.2. Minimum Description Length

MDL is a well known and powerful tool to estimate the best data model (among a class of candidates) and related parameters [21,22]. This principle allows for the selection of a good model for approximating the data with the least complexity. It is based on the rationale: good compression as good approximation, in agreement with the definition of Kolmogorov complexity [25]. In other words, given a finite-size data sample, the simplest model that well fits it is also the best one. The simplest formal way to implement MDL is the crude MDL. It selects a model $\bar{M}$ from a set $\mathcal{M}$ of candidates as it follows

$$\bar{M} = \underset{M \in \mathcal{M}}{argmin}\ L(M) + \lambda L(f|M) \tag{10}$$

where $L(M)$ is the cost (in terms of bits) required for coding the model M, $L(f|M)$ is the number of bits required for coding the data f given the model, while λ is a balancing param-

eter. In general, the better the model, the higher its cost, but the smaller the approximation error. That is why the selection of the best model is a trade off between complexity and good approximation. λ tuning represents a critical issue that is often solved empirically by properly adjusting the quantization step adopted for data coding or by properly selecting the coding algorithm [26]. Among the several applications of MDL-based strategy [27,28], it is worth mentioning the one recently presented in [29], where MDL was used for the selection of the number of components for PCA method [16]. As it is not trivial to practically define MDL, a linear regression model has been used as bound for its normalized version. In order to overcome this kind of problem, in this paper, we propose a different approach that simply consists of limiting the class of models to the one of the uniform sampling operator (of the feature matrix) and then using MDL for the selection of the best sampling step—in agreement with the standard sampling (stride) adopted in DNN architectures.

2.3. Mdl Based Selection of Wavelet Scattering Coefficients

In this paper, the normalized scattering coefficients $\tilde{S}$ in Equation (9) are properly modified in order to be considered as a distribution, and the corresponding entropy is used to define the coding lengths involved in the MDL functional.

To simplify the notation, the superscript ˜ will be omitted in the sequel. In addition, let

$$\mathbf{S}_{|p} = \mathbf{S} \odot \mathbf{T}_p \tag{11}$$

denote the subsampled scattering feature matrix along the frequency axis (row index), where $\odot$ is the Hadamard matrix product, p is the sampling step and $\mathbf{T}_p$ is the sampling matrix, and let

$$\mathbf{S}^c_{|p} = \mathbf{S} \odot \mathbf{T}^c_p \tag{12}$$

be its counterpart. Since the subsampling is odd, $\mathbf{S}_0$ is always preserved when subsampling $\mathbf{S}$, and the sampling matrix $\mathbf{T}_p$ is such that

$$T_p(i,j) = \begin{cases} 1 & i = hp,\ h \in \mathbf{N}, i \leq N, \quad j = 1, \ldots, N_t \\ 0 & \text{otherwise} \end{cases}, \tag{13}$$

while $\mathbf{T}^c_p = \mathbf{I} - \mathbf{T}_p$, with $\mathbf{I}$ as the all-ones matrix.

Now, let $\mathbf{P}$ be the $N \times N_t$ matrix, such that

$$\mathbf{P} = [P(i,j)]_{i=1,\ldots,N,\ j=1,\ldots,N_t} \tag{14}$$

with

$$P(i,j) = \frac{S^2(i,j)}{\|\mathbf{S}\|^2}. \tag{15}$$

The elements of $\mathbf{P}$ are positive; their value is less than one and defines a probability distribution.

The subsampled (by p) and rescaled distribution along the frequency axis is, therefore,

$$\mathbf{P}_{|p} = \frac{\|\mathbf{S}\|^2}{\|\mathbf{S}_{|p}\|^2}(\mathbf{P} \odot \mathbf{T}_p), \tag{16}$$

while its rescaled counterpart is

$$\mathbf{Q}_{|p} = \frac{\|\mathbf{S}\|^2}{\|\mathbf{S}^c_{|p}\|^2}(\mathbf{P} \odot \mathbf{T}^c_p). \tag{17}$$

Accordingly, the elements of $\mathbf{P}_{|p}$ and $\mathbf{Q}_{|p}$ define two distinct probability distributions. Therefore, according to Equation (10), the bits budget for the encoding error of the data, given the model ($L(f \mid M)$), is

$$L(f \mid p) = L(\mathbf{Q}_{|p}) = H(\mathbf{Q}_{|p}) \, \|\mathbf{S}^c_{|p}\|^2. \tag{18}$$

$L(\mathbf{Q}_{|p})$ is the entropy H of the data distributed as $\mathbf{Q}_{|p}$ multiplied by the amount of energy they convey. The latter is proportional to the number of elements, and it is necessary to express the cost in terms of bits. Accordingly, the cost of the model $L(M)$ should include both the cost of the sampling step p and the entropy of the data distributed as $\mathbf{P}_{|p}$, multiplied by their energy, i.e.,

$$L(\mathbf{P}_{|p}) = \lambda H(\mathbf{P}_{|p}) \, \|\mathbf{S}_{|p}\|^2 + 2\, log_2\lceil p \rceil + 1, \tag{19}$$

with λ as a proper balancing parameter. By setting $l(p) = 2log_2\lceil p \rceil + 1$ [21], the optimal sampling $\tilde{p}_f$ is then

$$\tilde{p}_f = \lfloor arg \min_p H(\mathbf{Q}_{|p}) \|\mathbf{S}^c_{|p}\|^2 + \lambda H(\mathbf{P}_{|p}) \|\mathbf{S}_{|p}\|^2 + l(p) \rceil, \tag{20}$$

where $\lfloor \cdot \rceil$ denotes the approximation to the nearest integer.

λ definition deserves some attention. By definition, WST layers do not have the same nature; all layers require high pass filtering operations before the application of the lowpass filter, except for $\mathbf{S}_0$. Dishomogeneity among layers is emphasized in the normalized scattering transform, because $\mathbf{S}_0$ does not have a parent. If this event does not influence $L(\mathbf{Q}_{|p})$, as it does not depend on $\mathbf{S}_0$, it is not so for $L(\mathbf{P}_{|p})$. Therefore, λ is required to compensate this imbalance. Specifically, it must depend on the probability that $\mathbf{S}_0$ is generated by the same source of the remaining normalized layers $\bar{\mathbf{S}} = \mathbf{S} - \mathbf{S}_0$, where $-$ denotes the difference between sets. To this end, the reciprocal relations between mean, standard deviation and energy of the two sources, $\mathbf{S}_0$ and $\bar{\mathbf{S}}$, are evaluated. In particular, a correction is needed whenever the contribution of $\mathbf{S}_0$ to the energy exceeds the one of $\bar{\mathbf{S}}$, its standard deviation is considerably smaller and the mean is very different. Hence, by denoting with μ_* and σ_*, respectively, the mean and the standard deviation (std) of $*$, and considering $\bar{\mathbf{S}}$ as a row vector,

STD if $\sigma_{\mathbf{S}_0} << \sigma_{\bar{\mathbf{S}}}$, then $\mathbf{S}_0$ resembles a uniform distribution. Hence, it satisfies the diffusivity property and its entropy dominates the one of the second source. A correction of the global entropy is then required accordingly, by measuring the probability $Pr\big(|S_0(j) - \mu_{\mathbf{S}_0}| \leq \sigma_{\bar{\mathbf{S}}}\big)$. The Chebyshev inequality [25] gives

$$Pr\big(|\mathbf{S}_0(j) - \mu_{\mathbf{S}_0}| \leq \sigma_{\bar{\mathbf{S}}}\big) > \left(1 - \frac{\sigma^2_{\mathbf{S}_0}}{\sigma^2_{\bar{\mathbf{S}}}}\right), \tag{21}$$

and the bound is not trivial whenever $\sigma^2_{\mathbf{S}_0} < \sigma^2_{\bar{\mathbf{S}}}$;

Mean if the previous condition holds and the mean values of $\mathbf{S}_0$ and $\bar{\mathbf{S}}$ are far apart, then the two sources are different. Since $\mu_{\mathbf{S}_0} - \mu_{\bar{\mathbf{S}}} = N(\mu_S - \mu_{\bar{S}})$, where N is the number of WST filters as defined in Equation (6), then

$$Pr\big(|\mu_{\mathbf{S}_0} - \mu_{\bar{\mathbf{S}}}| > \varepsilon\big) = Pr\Big(|\mu_{\mathbf{S}} - \mu_{\bar{\mathbf{S}}}| > \frac{\varepsilon}{N}\Big) \leq \frac{N^2\sigma^2_{\bar{\mathbf{S}}}}{\varepsilon^2} \tag{22}$$

where the Chebyshev inequality [25] extended to the sample mean has been applied. ε has been set equal to $\frac{N \cdot N_t}{\sqrt{12}}$ and denotes the std of a diffusive WST;

Energy to check if the contribution to the energy of $\mathbf{S}_0$ is greater than the one of $\bar{\mathbf{S}}$, $Pr\Big(|S_0(j)|^2 \geq \frac{\|\bar{S}\|^2}{N_t}\Big)$ is estimated. By denoting with n the number of WST coeffi-

cients, i.e., $n = N \cdot N_t$, the Markov inequality [25], extended to the square root function, gives

$$Pr\left(|\mathbf{S}_0(j)|^2 \geq \frac{\|\bar{\mathbf{S}}\|_2^2}{N_t}\right) \leq \frac{\frac{\|\bar{\mathbf{S}}\|_1}{n-N_t}}{\frac{\|\mathbf{S}_0\|_2}{\sqrt{N_t}}} \leq \frac{\frac{\|\bar{\mathbf{S}}\|_2}{\sqrt{n-N_t}}}{\frac{\|\mathbf{S}_0\|_2}{\sqrt{N_t}}}. \tag{23}$$

The equivalence between compatible norms has been used to obtain the rightmost bound that is not trivial if $\frac{\|\bar{\mathbf{S}}\|_2 \sqrt{N_t}}{\|\mathbf{S}_0\|_2 \sqrt{n-N_t}} \leq 1$;

By combining Equations (21)–(23), λ can then be defined as

$$\lambda = \begin{cases} 1 & \sigma^2_{\mathbf{S}_0} \geq \sigma^2_{\bar{\mathbf{S}}} \\ \left(1 - \frac{\sigma^2_{\mathbf{S}_0}}{\sigma^2_{\bar{\mathbf{S}}}}\right) \frac{12\sigma^2_{\bar{\mathbf{S}}}}{N_t^2} \min\left(1, \frac{\frac{\|\bar{\mathbf{S}}\|_2}{\sqrt{n-N_t}}}{\frac{\|\mathbf{S}_0\|_2}{\sqrt{N_t}}}\right) & \sigma^2_{\mathbf{S}_0} < \sigma^2_{\bar{\mathbf{S}}} \end{cases}. \tag{24}$$

This makes the proposed method completely automatic.

2.4. The Algorithm

Let $\mathcal{T}$ be the training set. The algorithm consists of the following steps.

1. For each signal f in $D \subset \mathcal{T}$, fixed number of layers (m) and Q factors:
   - Compute the normalized WST (feature matrix) of f as in Equation (9) and the distribution matrix $\mathbf{P}$ as in Equation (14).
   - For each sampling step $p = 1, 2, 3...$, compute $\tilde{p}_f$ by minimizing the MDL functional as in Equation (20).
2. Set the optimal sampling step $\tilde{p} = \left\lfloor \frac{1}{|D|} \sum_{f \in D} \tilde{p}_f \right\rceil$, with $|D|$ as the number of signals in D. It is the average of the sampling steps estimated in step 1 for each f.
3. Apply SVM to estimate the classification model by using the sampled distribution matrix $\mathbf{S}_{|\tilde{p}}$ of each signal in $\mathcal{T}$ as input.

3. Results

The proposed MDL-based selection strategy has been applied to different datasets of sound signals. This section refers to three datasets: GTZAN [30], PhysioNet (ECG) [31] and the Free Spoken Digits Database (FSDD) [32]. The GTZAN dataset is widely used for comparative studies in music genre classification. It includes 10 genres, each containing 100 clips of 30 s sampled at 22,050 Hz. The second dataset consists of 162 ECG recordings obtained from three groups of people with: cardiac arrhythmia (96), congestive heart failure (30) and normal sinus rhythms (36). The Spoken Digit Dataset consists of recordings of spoken digits in 'wav' files sampled at 8 kHz. It is an open dataset that grows over time. The one used in the tests (downloaded on 17 May 2021) consists of 3000 recordings of digits zero through nine, pronounced by six English speakers. Equal-length signals, three layers ($m = 2$) WST with different Q factors and a polynomial kernel-based SVM classifier, have been used in all tests. The percentage of each class for training and test sets for each dataset has been, respectively, 80–20 (GTZAN), 70–30 (PhysioNet) and 80–20 (FSDD).

Results have been evaluated in terms of classification accuracy and with respect to the goals of the paper:

(i) Preservation or improvement of the classification accuracy provided by the full WST feature matrix for fixed Q factors;

(ii) Reduction in the learning time in terms of reduced number of weights to learn;

(iii) Definition of an automatic procedure.

They have been compared with PCA-based scattering features selection and WST layer-selection methods, as in the seminal papers [1,11].

Regarding points (i) and (ii), Table 1 refers to the Physionet dataset and five couples of Q factors. In this case, normalized WST (*3rd column*) easily reaches the classification task, independently of WST parameters. On the other hand, a reduced number of scattering coefficients (*fourth column*) allows us to reach the classification task too, while reducing the complexity of classification algorithm, as a lower number of weights has to be estimated by the classifier. The gain is not negligible, as sampling reduces the number of features up to 25% ($\tilde{p} = 4$) of the full matrix. Table 1 also compares the results achieved by the proposed uniform sampling to the ones achieved using a lower number of layers, as shown in [1]. As can be observed in the last three columns of the table, the use of a smaller number of layers cannot guarantee the same results, in terms of accuracy and/or number of features, of the suitably sampled WST feature matrix. On the one hand, the second layer allows for high classification accuracy but retains a large number of features; on the other hand, the first two layers ($0th$ and $1st$) retain a smaller number of features: not enough to exactly assess cardiac conditions.

Table 1. Physionet dataset: Classification accuracy (%) for different couples of Q factors. The feature matrix consists of the logarithm of: WST (*2nd col*); normalized WST (*3rd col*); the uniformly sampled feature matrix (normalized WST) using the estimated sampling step $\tilde{p}$, as in Equation (20) (*4th col*); normalized WST coefficients which, respectively, belong to the 0th and 1st layer, only the 1st layer, only the 2nd layer (*last three cols*). The number of features for each time t is in round brackets, while the value of $\tilde{p}$ is in square brackets. Best results are in bold.

Q_1, Q_2	$log(\mathbf{S})$	$log\tilde{\mathbf{S}}$	$\tilde{p}$	$log\tilde{\mathbf{S}}_0, log\tilde{\mathbf{S}}_1$	$log\tilde{\mathbf{S}}_1$	$log\tilde{\mathbf{S}}_2$
3, 2	95.92 (395)	100 (395)	**100** [3] (132)	95.9 (35)	89.8 (34)	95.9 (361)
4, 2	95.92 (483)	100 (483)	**100** [3] (161)	98.0 (45)	93.9 (44)	100 (438)
4, 3	97.96 (721)	100 (721)	**100** [4] (181)	98.0 (45)	93.9 (44)	100 (676)
8, 1	97.96 (409)	100 (409)	**100** [2] (205)	98.0 (84)	98.0 (83)	93.9 (325)
8, 3	97.9 (1221)	100 (1221)	**100** [4] (306)	97.9 (84)	97.9 (83)	100 (1137)

Regarding points (i) and (iii), results presented in Table 2 aim to show that uniform sampling can provide non-negligible gain in terms of accuracy and that the proposed method is able to correctly estimate the required sampling step. To this aim, some representative results obtained using different couples of Q factors for the three datasets are shown. The same results are compared to those achieved when using PCA to reduce the dimension of the WST feature matrix (*last three columns*), as shown in [1,17]. As can be observed, a reduced number of scattering coefficients (sampling $p = 2, 3, 4$) can provide higher classification accuracy than using the full feature matrix (sampling $p = 1$). In addition, the proposed MDL-based procedure is able to correctly guess the sampling $\tilde{p}$, providing the highest classification accuracy in most cases. In addition, if more than one sampling step guarantees the best classification accuracy, the proposed method selects the one that provides the highest (or nearly the highest) data reduction in terms of number of retained scattering samples. With regard to this point, it is worth observing that sometimes the method can fail to predict the optimal sampling, as the latter is defined as the average of the optimal sampling steps that are estimated from each signal independently. More accurate estimations can be obtained by refining the averaging adopted in step 2 of the algorithm, e.g., by discarding eventual outliers or unacceptable solutions, and this will be the topic of future work. Regardless, without applying any correction, for the three datasets and

several couples of Q factors, the measured success rate for this preliminary version of the method was about 93%.

Table 2. *1st col:* Dataset; *2nd col:* WST Q factors; *Cols 3–6:* Classification accuracy (%) for different sampling steps—the number of samples is in the brackets; *7th col:* Optimal sampling step selected by the proposed method; *Cols 8–9:* PCA-based classification: classification accuracy (%) by retaining those principal components expressing the 95% and the 99% of the total variance of the feature matrix (the number of components is in the brackets); *Last col:* PCA-based classification: classification accuracy (%) by retaining a number of principal components equal to the frequency samples (in the brackets) retained when the sampling step $\tilde{p}$ is applied. Best results are in bold.

Dataset	Q_1, Q_2	**Sampling** p				$\tilde{p}$	**PCA**		
		1	**2**	**3**	**4**		**95%**	**99%**	$\tilde{p}$
FSDD	4, 2	96.2 (265)	**96.8** (133)	96.2 (89)	95.2 (67)	2	86.7 (13)	94.7 (226)	95.0 (133)
FSDD	4, 3	95.3 (367)	**96.5** (184)	94.7 (123)	93.3 (92)	2	71.0 (8)	91.7 (23)	93 (184)
FSDD	5, 2	96.7 (311)	96.8 (156)	96.3 (104)	**97.2** (78)	4	88.5 (14)	95.3 (33)	96.7 (78)
GZTAN	4, 2	87.5 (500)	**90.0** (250)	88.5 (167)	82.5 (125)	2	88.5 (110)	86.5 (87)	87 (250)
GZTAN	4, 3	89.0 (595)	**90.0** (298)	88.5 (199)	88.0 (149)	2	87.0 (176)	87.5 (336)	88 (298)
GZTAN	8, 1	85 (341)	86.5 (171)	**88** (114)	85 (86)	3	89 (87)	87.5 (201)	89.5 (114)
Physionet	2, 1	95.9 (143)	95.9 (72)	**98.0** (48)	91.8 (38)	3	93.9 (19)	93.9 (55)	98.0 (48)
Physionet	4, 1	**100** (241)	98.0 (121)	98.0 (81)	95.9 (61)	3	98.0 (27)	100 (86)	98.0 (241)
Physionet	6, 2	**100** (659)	**100** (330)	**100** (220)	**100** (165)	3	100 (45)	98.0 (179)	100 (165)

With regard to PCA-based feature reduction, two different criteria for the selection of the number of components have been adopted. The former is the standard selection of those components retaining a predefined percentage of variance (*cols 7–8*); the latter selects the first L principal components, with L equal to the first dimension of the sampled WST feature matrix that is obtained using $\tilde{p}$ as sampling step (*last col*). Table 2 emphasizes two interesting aspects. The first one is that PCA + SVM does not provide the best classification accuracy if the number of principal components is estimated by retaining the principal components conveying the highest percentage of variance (*cols 7–8*). A criterion for selecting the best percentage of preserved variance is then required, either for maximizing classification accuracy or for minimizing the number of components providing the same accuracy. This further gives evidence of the need for an automatic and effective selection of significant components. The second one is that for a fixed number of components, i.e., the one corresponding to the optimal sampling step, the proposed method provides classification accuracies that are comparable to—or even better than—the one provided by PCA (*last col*)—this holds for all Q factors pairs in Table 2, except for the $7th$ row. As a result, the selection of some samples from each frequency band can represent a robust approach. In addition, it is less time consuming, and thus is computationally advantageous.

To further evaluate the proposed approach, some feature ranking methods have been adopted for the selection of significant scattering coefficients. Table 3 shows some results achieved on FSDD dataset. They refer to the minimum redundancy maximum relevance (MRMR) algorithm. It is a filter-type feature selection method that ranks the features by using mutual information [33]. The table shows the classification rates achieved by using

the first most significant ranked features that are selected so that the sum of their ranking scores equals a predefined percentage of the overall ranking score. As can be observed, the number of features whose global ranking exceeds 90% is higher than the one given by the optimal uniform sampling step that is estimated by the proposed method. In addition, the selected features do not allow us to reach the same classification rates. This confirms the proposed approach as a reliable and effective feature selection method, even though it is restricted to the uniform sampling procedure. Table 4 refers to FSDD and Physionet datasets and reports the classification rates achieved using a sequential selection criterion (wrapper-type feature selection method). In this case, features are selected on the basis of the multiclass error-correcting output codes (ECOC) model using SVM binary learners.

Table 3. FSDD Dataset; *1st col:* WST Q factors; *2nd col:* Classification accuracy (%) for the optimal sampling step; *Cols 3–10:* MRMR ranking-based feature selection: classification accuracy (%) by retaining the first most significant features, whose global ranking is a predefined percentage (respectively, 5%, 10%, 20%, 30%, 40%, 50%, 75%, 90%). The number of features is in the brackets. Best results are in bold.

Q_1, Q_2	Sampling $\tilde{p}$	MRMR							
		5%	**10%**	**20%**	**30%**	**40%**	**50%**	**75%**	**90%**
4, 2	**96.8** (133)	95.2 (43)	94.8 (72)	95.3 (106)	95.2 (134)	95.7 (156)	95.7 (156)	95.8 (183)	96.0 (225)
4, 3	**96.5** (184)	94.7 (37)	95.0 (73)	95.0 (106)	95.3 (150)	95.8 (193)	95.7 (214)	95.3 (271)	95.0 (319)
5, 2	**97.2** (78)	95.7 (43)	96.8 (72)	96.5 (108)	96.0 (128)	96.3 (154)	96.2 (181)	96.5 (208)	96.8 (261)

Table 4. *1st col:* Dataset; *2nd col:* WST Q factors; *3rd col:* Classification accuracy (%) for the optimal sampling step—and the one estimated by the proposed method if different from the expected one; *4th col:* classification accuracy for the sequential feature selection method (SFS). The number of selected features is in the brackets. Best results are in bold.

Dataset	Q_1, Q_2	**Sampling** $\tilde{p}$	**SFS**
Physionet	2, 1	**98.0** (48)	87.8 (24)
Physionet	4, 1	**100**—98.0 (241)—(81)	98.0 (18)
Physionet	6, 2	**100**—100 (165)—(220)	91.8 (17)
FSDD	4, 2	**96.8** (133)	95.5 (24)
FSDD	4, 3	**96.5** (184)	94.5 (20)

As can be observed, the sequential feature selection (SFS) is shown to be too conservative. In fact, it selects a very small number of features, and reaches satisfying classification rates only for some couples of Q-factors. Moreover, it requires significant computational effort: the required cpu time is at least 10 times greater than the one required by the proposed selection method when running on the same machine. It is also worth observing that, even when the proposed method is not able to select exactly the best sampling step, it allows us to reach the highest classification rates, as in the case of the couple $(6, 2)$ in the Physionet dataset, or the same performance of the SFS method but requiring a considerably lower computing time (Physionet dataset, couple $(4, 1)$).

4. Conclusions

In this paper, the first study concerning the best selection of scattering coefficients for sound signals classification was presented. Uniform sampling was adopted and a MDL-based model selection procedure was defined. The main goal was to establish to what extent an automatic procedure for dimensionality reduction, although simple, can contribute to improving signal classification tasks in terms of both computing time and accuracy. One of the main efforts required weighting a MDL functional through a data-dependent parameter; the latter plays a key role in the proposed approach, as any user's intervention is required, as well as any a priori information concerning the signal type. The use of WST for the classification of sound signals allows us to exploit the benefits coming from a CNN architecture, while both reducing training time, as WST emphasizes distinctive time-frequency structures, and tuning the automatic procedure. In addition, the latter requires very little computational effort, as it consists of sampling and energy computation. Future study will focus on refining the definition of MDL functional used for model selection as well, as its generalization to a wider class of models.

Author Contributions: Conceptualization, V.B. and D.V.; methodology, V.B. and D.V.; software, M.L.C.; validation, V.B., M.L.C. and D.V.; writing—original draft preparation, V.B.; writing—review and editing, V.B., M.L.C. and D.V. 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: This research was partially funded by the Italian national research group GNCS (INdAM). This research has been accomplished within RITA (Research ITalian network on Approximation).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

WST	Wavelet Scattering Transform;
SVM	Support Vector Machine;
CNN	Convolutional Neural Network;
DNN	Deep Neural Network;
MDL	Minimum Description Length;
PCA	Principal Component Analysis;
MDS	Multidimensional Scaling.

References

1. Anden, J.; Mallat, S. Deep Scattering Spectrum. *IEEE Trans. Signal Process.* **2014**, *62*, 4114–4128. [CrossRef]
2. Anden, J.; Lostanlen, V.; Mallat, S. Joint Time–Frequency Scattering. *IEEE Trans. Signal Process.* **2019**, *67*, 3704–3718. [CrossRef]
3. Bruna, J.; Mallat, S. Invariant Scattering Convolution Networks. *IEEE Trans. PAMI* **2013**, *35*, 1872–1886. [CrossRef] [PubMed]
4. Chin, C.; Zhang, J. Wavelet Scattering Transform for Multiclass Support Vector Machines in Audio Devices Classification System. In Proceedings of the IEEE/ASME AIM 2021, Delft, The Netherlands, 12–16 July 2021.
5. Ghezaiel, W.; Brun, L.; Lezoray, O. Wavelet Scattering Transform and CNN for Closed Set Speaker Identification. In Proceedings of the IEEE MMSP 2020, Virtual, 21–24 September 2020.
6. Hajihashemi, V.; Gharahbagh, A.A.; Cruz, P.M.; Ferreira, M.C.; Machado, J.J.M.; Tavares, J.M.R.S. Binaural Acoustic Scene Classification Using Wavelet Scattering, Parallel Ensemble Classifiers and Nonlinear Fusion. *Sensors* **2022**, *22*, 1535. [CrossRef] [PubMed]
7. Kanalici, E.; Bilgin, G. Music Genre Classification via Sequential Wavelet Scattering Feature Learning. In Proceedings of the KSEM 2019, Athens, Greece, 28–30 August 2019.

8. Oyallon, E.; Belilovsky, E.; Zagoruyko, S.; Valko, M. Compressing the Input for CNNs with the First-Order Scattering Transform. In Proceedings of the ECCV 2018, Munich, Germany, 8–14 September 2018.
9. Song, G.; Wang, Z.; Han, F.; Ding, S. Transfer Learning for Music Genre Classification. In Proceedings of the ICIS 2017, South, Korea, 10–13 December 2017.
10. Baseer Buriro, A.; Ahmed, B.; Baloch, G.; Ahmed, J.; Shoorangiz, R.; Weddell, S.J.; Jones, R.D. Classification of alcoholic EEG signals using wavelet scattering transform-based features. *Comput. Biol. Med.* **2021**, *139*, 104969. [CrossRef]
11. Anden, J.; Mallat, S. Multiscale scattering for audio classification. In Proceedings of the ISMIR 2011, Miami, FL, USA, 24–28 October 2011.
12. Lostanlen, V.; Cohen-Hadria, A.; Pablo Bello, J. One or Two Frequencies? The Scattering Transform Answers. In Proceedings of the EUSIPCO, Dublin, Ireland, 23–27 August 2021.
13. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **2014**, *40*, 16–28. [CrossRef]
14. Cox, M.; Cox, T. Multidimensional Scaling. In *Handbook of Data Visualization*; Springer Handbooks Comp.Statistics; Springer: Berlin/Heidelberg, Germany, 2008.
15. Ferreira, A.J.; Figueiredo, M.A.T. Efficient feature selection filters for high-dimensional data. *Pattern Recognit. Lett.* **2012**, *33*, 1794–1804. [CrossRef]
16. Jolliffe, I.; Cadima, J. Principal component analysis: A review and recent developments. *Philosphiocal Trans. A* **2016**, *374* . [CrossRef] [PubMed]
17. Li, J.; Ke, L.; Du, Q.; Ding, X.; Chen, X.; Wang, D. Heart Sound Signal Classification Algorithm: A Combination of Wavelet Scattering Transform and Twin Support Vector Machine. *IEEE Access* **2019**, *7*, 179339–179348. [CrossRef]
18. Liu, Z.; Yao, G.; Zhang, Q.; Zhang, J.; Zeng, X. Wavelet Scattering Transform for ECG Beat Classification. *Comp. Math. Methods Med.* **2020**, *2020*. [CrossRef] [PubMed]
19. Rodriguez-Algarra, F.; Sturm, B.L. Re-evaluating the scattering transform. In Proceedings of the ISMIR 2015, Malaga, Spain, 26–30 October 2015.
20. Solorio-Fernández, S.; Carrasco-Ochoa, J.A.; Martínez-Trinidad, J.F. A review of unsupervised feature selection methods. *Artif. Intell. Rev.* **2020**, *53*, 907–948. [CrossRef]
21. Grunwald, P.D.; Grunwald, A. *The Minimum Description Length Principle*; MIT Press: Cambridge, MA, USA , 2007.
22. Hu, B.; Rakthanmanon, T.; Hao, Y.; Evans, S.; Leonardi, S.; Keogh, E. Using the minimum description length to discover the intrinsic cardinality and dimansionality series. *Data Min. Knowl. Discov.* **2015**, *29*, 358–399. [CrossRef]
23. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*; Cambridge University Press: Cambrigde, UK, 2000.
24. Bruna, J.; Mallat, S. Classification with Scattering Operators. In Proceedings of the IEEE CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011.
25. Cover, T.M. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 1999.
26. Grunwald, P. *Minimum Description Length Tutorial*; Advances in MDL: Theory and Applications; MIT Press: Cambridge, MA, USA , 2005; pp. 23–80.
27. Bruni, V.; Vitulano, D. An entropy based approach for SSIM speed up. *Signal Process.* **2017**, *135*, 198–209. [CrossRef]
28. Bruni, V.; Cardinali, M.L.; Vitulano, D. A Short Review on Minimum Description Length: An Application to Dimension Reduction in PCA. *Entropy* **2022**, *24*, 269. [CrossRef] [PubMed]
29. Tavory, A. *Determining Principal Component Cardinality through the Principle of Minimum Description Length*; LNCS; Springer: Cham, Switzerland, 2019; Volume 11943, pp. 655–666.
30. Tzanetakis, G.; Cook, P. Music genre classification of audio signals. *IEEE Trans. Speech Audio Process.* **2002**, *10*, 293–302. [CrossRef]
31. Goldberger, A.L.; Amaral, L.; Glass, L.; Hausdorff, J.M.; Ivanov, P.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.K.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation* **2000**, *101*, 215–220. [CrossRef] [PubMed]
32. Free Spoken Digit Dataset (FSDD). Available online: https://github.com/Jakobovski/free-spoken-digit-dataset (accessed on 17 May 2021).
33. Ding, C.; Peng, H. Minimum redundancy feature selection from microarray gene expression data. *J. Bioinform. Comput. Biol.* **2005**, *3*, 185–205. [CrossRef] [PubMed]