The Importance of Multiple Temporal Scales in Motion Recognition: from Shallow to Deep Multi Scale Models

Vincenzo D'Amato, Luca Oneto,
Antonio Camurri, Davide Anguita
University of Genoa
Via Opera Pia 11a, 16145, Genova, Italy
{vincenzo.damato,luca.oneto}@unige.it
{antonio.camurri,davide.anguita}@unige.it

Zinat Zarandi, Luciano Fadiga,
Alessandro D'Ausilio, Thierry Pozzo
University of Ferrara
Via Fossato di Mortara 19, 44121, Ferrara, Italy
{zinat.zarandi,luciano.fadiga}@iit.it
{alessandro.dausilio,thierry.pozzo}@iit.it

2022 International Joint Conference on Neural Networks (IJCNN) | 978-1-7281-8671-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/IJCNN55064.2022.9892519

Abstract—**Studying human motion requires modelling its multiple temporal scale nature to fully describe its complexity since different muscles are activated and coordinated by the brain at different temporal scales in a complex cognitive process. Nevertheless, current approaches are not able to address this requirement properly, and are based on oversimplified models with obvious limitations. Data-driven methods represent a viable tool to address these limitations. Nevertheless, shallow data-driven models, while achieving reasonably good recognition performance, require to handcraft features based on domain-specific knowledge which, in this cases, is limited and does no allow to properly model motion- and subject-specific temporal scales.**

In this work, we propose a new deep multiple temporal scale data-driven model, based on Temporal Convolutional Networks, able to automatically learn features from the data at different temporal scales. Our proposal focuses first on over-performing state-of-the-art shallows and deep models in terms of recognition performance. Then, thanks to the use of feature ranking for shallow models and an attention map for deep models, we will give insights on what the different architectures actually learned from the data.

We designed, collected data, and tested our proposal in custom experiment of motion recognition: detecting the person who draw a particular shape (i.e., an ellipse) on a graphics tablet, collecting data about his/her movement (e.g., pressure and speed) in different extrapolating scenarios (e.g., training with data collected from one hand and testing the model on the other one). Collected data regarding our experiment and code of the methods are also made freely available to the research community.

Results, both in terms of accuracy and insight on the cognitive problem, support the proposal and support the use of the proposed technique as a support tool for better understanding the human movements and its multiple temporal scale nature.

Index Terms—**Motion Recognition, Multiple Temporal Scales, Shallow Learning, Deep Learning, Feature Engineering, Feature Learning, Attention Maps, Open Data, Open Implementation**

I. Introduction

The study of human movement is a broad and multidisciplinary research field, including disciplines such as cognitive neuroscience, experimental psychology, biomechanics, interaction design, artificial intelligence, and theories from the arts [1]–[3]. According to [4], three main research areas are surveillance (e.g., check for critical events such as fall detection in frail elderly), control (e.g., improve the mobility of a patient), and analysis (e.g., understand the quality of full-body movement in sports or in expressive emotional communication). Studies in human movement find a large number of applications [2], including physical rehabilitation, sport scoring and skill assessment [5], [6] and applications involving the full-body expression of emotions and non-verbal social signals. Each research field tries to address and study human movements from its own perspective [3], [7]–[11]. For instance, experimental psychology and cognitive neuroscience provide theoretical frameworks and cognitive models [11], [12], while computational methods (e.g., data mining and machine learning) can leverage on data collected by specific sensors (e.g., video, motion capture, and inertial sensors) to provide insights on movement qualities [3], [10]. Recent studies [13]–[16] show how human movements are hierarchically nested: i.e., a layered movement structure involves muscles of different size, responsiveness, and precision, with a complex non-linearly stratified temporal dimension where every muscle has its own temporal scale. Even the simplest actions (e.g., point a finger toward an object) are composed by a set of sub-actions involving different parts of the body (e.g., pointing a finger toward an object starts from the movement of the entire body, followed by the upper part, then the shoulder, the arm, the finger, etc.), which cooperate to create a resulting smooth, effective, and expressive movement in a complex multiple temporal scale cognitive task. The human body can be described as a kinematic tree, consisting of joints connected to each other. As a first approximation, we can state that larger muscles are slower and characterised by a slower perceptual response over time with respect to the smaller muscles. Nevertheless, some movements of larger muscles can be fast: for example, the small corrections to keep us in balance to compensate for a loss of balance, to avoid the risk of fall. Note also that the multiple temporal scales nature of the human movements, also characterises how humans perceive other people's movements [7], [17]. Human beings are capable of

978-1-7281-8671-9/22/$31.00 ©2022 IEEE

understanding and predicting the movements of other humans even from a limited number of moving points [18]. This skill depends on the ability of humans to create relations between different temporal and spatial layers using forward/feedback connections. This process is driven by the brain and the body as time keepers that coordinate different internal, mental, and physiological clocks. Nevertheless, it is worth noting that recent studies [19] demonstrate that the information contained in these limited number of moving points does not depend only on the activity performed but also on more complex cognitive and affective phenomena. For example, Meeren et al. [20] consider the relation stimulus, defined as temporal dynamics of the feedback, and affective qualities.

Although it is nowadays well known that human movement and perception of human movement are characterised by multiple temporal scales [7], [8], [13]–[17], [20], very few works in the literature are focused on studying this particular property. For instance, Ihlen et al. [21] provided a quantitative support for studying the multiple temporal scales in human action and perception using wavelet-based multifractal analysis in the response series of four cognitive tasks (simple response, word naming, choice decision and interval estimation). Camurri et al. [22] demonstrate that computational models of expressive qualities should operate at different temporal scales starting from a previous research on human perception and dance theories [23]. Authors of [22] propose a framework where features are computed at different levels, i.e., low-level features (e.g., velocity) are computed instantaneously while higher ones (e.g., impulsiveness) are computed on a larger temporal scale. In image recognition tasks like object detection, semantic segmentation and action recognition, Temporal Convolutional Networks (TCNs) with dilated convolutions [24]–[26] have been widely adopted to increase receptive field sizes without increasing model complexity. Indeed, by applying dilated convolutions with different filter sizes, multiple temporal scales can be efficiently captured and the use of this mathematical operation can handle larger temporal context efficiently. A recent research, carried out in the European FET PROACTIVE Project EnTimeMent[1], focuses its attention on addressing the importance of multiple temporal scales in movement analysis and prediction. Inside EnTimeMent, Beyan et al. [27] propose an approach able to model the dynamics of full-body movement data represented on multiple temporal scales where features are processed by two independent and parallel shallow TCNs.

In this work, we investigate how data-driven methods can represent research frontier to validate some results from cognitive neuroscience in understanding human movement. In particular, we will first show how shallow data-driven models [28], namely models which require to handcraft features based on domain-specific knowledge, can achieve good recognition performance, but they are hardly able to model the multiple temporal scales characterising human movement. Consequently, we propose a new deep multiple temporal scale data-driven model [29] able to automatically learn features from the data at different temporal scales and over-perform state-of-the-art shallow models in terms of recognition performance. For what concerns the shallow model we rely on Random Forest (RF) [30] and state-of-the-art signal processing techniques for feature engineering [10], [31], customised for the particular problem under exam. For what concerns the deep model, we first adopt a classical Long-Short Term Memory (LSTM) [14], [32], [33] Network architecture, showing its inability to over-perform RF plus carefully handcrafted feature, since it is not able to adequately capture the multiple temporal scales of the phenomena. Then, we propose a new architecture based on TCN [24]–[26], [34]–[37] capable to capture and learn the multiple temporal scale nature of the phenomena under investigation. In order to provide more insights on the cognitive task understanding what the data-driven models actually learned from the data we will leverage on feature ranking [38]–[40] for shallow models and an attention map [41] for deep models.

To prove the effectiveness of our proposal, we designed, collected data, and tested it in the following experiment on movement recognition: detecting the person who draws a shape (an ellipse) on a graphics tablet. We collected data on his/her movement (position, pressure, speed) in different extrapolating scenarios (e.g., training with data collected from one hand and testing the model on the other one). This experiment is inspired by the work of Scocchia et al. [42], where the authors exploit simplified models to measure the different perception of people in observing a moving dot along an elliptical trajectory. A publicly available dataset of our experiment is available for the scientific community together with the code developed to generate all the results presented in this paper[2].

The rest of the paper is organised as follows. Section II describes our experiment on human movement and details the related collected data. Section III presents state-of-the-art methods and the newly proposed ones that we employed in this work. Section IV reports the results in applying the methods described in Section III on the data collected in Section II. Section V concludes the paper.

II. Experiment of Motion Recognition and Related Data

Human motion is characterised by geometric and kinematic patterns that can be explained by a limited number of laws of motion. In particular, the two-thirds power law is able to model the relation between velocity and curvature typical of human movement [43], [44]: when velocity decreases the curvature increases and vice versa. The two-thirds power law can describe a variety of movement tasks and muscle districts involved in such tasks, including planar drawing movements [43], [44] or the perception of movements [42], [44]. Unfortunately, the two-thirds power law is a too simple model to be exploited in practice. The definition of a richer

[1] https://entimement.dibris.unige.it/

[2] https://github.com/lucaoneto/IJCNN2022_Ellipses

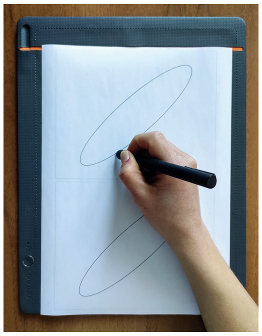


Fig. 1: Example of data acquisition with the graphics tablet[3].

model, capable of explaining the differences between one individual to another, is necessary. This challenge is very difficult in real-world scenarios. We designed an experiment in a simplified scenario, inspired by the work of Scocchia et al. [42], who explore the different perception of individuals in observing a moving dot along an elliptical trajectory. We designed and collected data on different individuals who were all asked to draw an ellipse on a graphics tablet: the goal is to detect each individual person from the details on how s/he draws the ellipse.

We collected data using a graphics tablet[3], under an ordinary lighting condition and vertically positioned respect to the sitting participant (see Figure 1). We collected data about 14 right-handed subjects who were asked to draw several times an ellipse. We varied the hand (left and right) and the drawing speed (slow, medium and fast according to the sensibility of the subject). The direction of the ellipses is different based on the hands (clockwise for the right hand and anticlockwise for the left hand) in order to make the drawing phase more natural and instinctive. For each combination we asked to repeat the draw 10 times where we discarded the first 2 and the last one to avoid a border effect. Each session of the experiment is repeated 10 times. The resulting dataset consists of the following recordings: 14 participants, 2 hands, 3 speeds, 7 kept ellipses, 10 repetition of the experiment. From the total of 5880 recordings, we selected ≈5663, since some of them were corrupted. For each recording we collected a time series reporting position of the pencil on the graphics tablet $(x(t), y(t))$ and the pressure $p(t)$ with a sampling rate of 0.01 seconds. From the position we compute the angular velocity $v(t)$ and the radius of curvature $r(t)$, which together with the $p(t)$ are the most representative information on the movement.

III. Methods

In this section, after a brief section of preliminaries (Section III-A) we will first (Section III-B) present the state-of-the-art shallow model based on RF and the related feature engineering process that we customised for the particular problem under exam. Moreover, in order to understand how the algorithm actually exploits the derived features to make a prediction a feature ranking phase is also performed. Then (Section III-C) we will present the deep state-of-the-art counterpart based on TCN capable to over-perform both state-of-the-art shallow and deep models in our specific problem. For TCN, in order to understand what the network actually learned from the data, we will rely on the attention mechanism to derive some attention maps. Finally, details on the model selection and error estimation phases are reported (see Section III-D), to support the statistical validity of the results. Different scenarios of extrapolation will be considered as detailed in Section III-E.

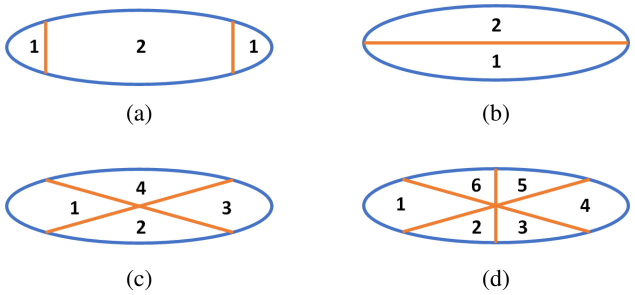
A. Preliminaries

As described in Section II the problem that we want to solve is a now-classical multiclass classification problem [28] whose output is what subject has drawn a particular ellipse and whose inputs are the time series $v(t)$, $r(t)$, and $p(t)$. In particular, in multiclass classification an input space $\mathcal{X}$ (the three time series $v(t)$, $r(t)$, and $p(t)$ collected when writing a particular ellipse) and an output space $\mathcal{Y}$ (the subject who wrote the ellipse) are available. Let $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$, where $X_i \in \mathcal{X}$ and $Y_i \in \mathcal{Y}$ $\forall i \in \{1, \dots, n\}$, be a sequence of n samples drawn from $\mathcal{X} \times \mathcal{Y}$. Let us consider a model (function) $f : \mathcal{X} \to \mathcal{Y}$ chosen from a set $\mathcal{F}$ of possible hypotheses. For shallow models $\mathcal{X}$ is first mapped, by means of an handcrafted feature engineering phase, into a vector $\phi(X) \in \mathbb{R}^d$, named representation, able to well represent $\mathcal{X}$ while discarding the not useful information. For deep models $\mathcal{X}$ is mapped into a representation $\phi(X)$ by means of an architecture able to automatically learn the best mapping. An algorithm $\mathscr{A}_{\mathcal{H}} : \mathcal{D}_n \times \mathcal{F} \to f$ characterised by its hyperparameters (e.g., deep of a tree in a RF or size of the convolution filters in a TCN) $\mathcal{H}$ selects a model inside a set of possible ones based on $\mathcal{D}_n$. The error of f in approximating $\mathbb{P}\{Y|X\}$ is measured by a prescribed metric $M : \mathcal{F} \to \mathbb{R}$. For what concerns the $M(f)$ many different metrics are available in literature [28]. In this work we will exploit the percentage of accuracy (ACC), the precision (PRE), the recall (REC), and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC). In particular, since our problem is a multiclass classification problem and consequently REC, PRE, and ROC-AUC cannot be directly defined, we will use the One Versus One (OVO) schema which compares every unique pairwise combination of classes and then takes the average behaviour. In order to tune the performance of the $\mathscr{A}_{\mathcal{H}}$, namely to select the best set of hyperparameters, and to estimate the performance of the final model according to the desired metrics, a Model Selection (MS) and Error Estimation (EE) phase need to be performed [45]. Moreover, in order to give some insights on what the algorithms actually learned from the data it is required to provide some explainability properties of the learned models [38]–[41]. For shallow models, feature ranking, namely how much the handcrafted features actually contribute to the prediction, is surely one of the most effective tools. For deep models, exploiting the attention mechanism,

[3] Wacom Bamboo slate; temporal resolution: 200 samples/s; resolution: 1748 by 2551; Active area: 210 × 297 mm

Function	Description
mean	Mean value
var	Variance
mad	Median absolute value
max	Largest value in array
min	Smallest value in array
sma	Signal magnitude area
energy	Average sum of squares
iqr	Interquartile range
entropy	Signal Entropy
correlation	Correlation coefficient between series
kurtosis	Signal Kurtosis
skewness	Signal Skewness
maxFreqInd	Largest frequency component
argMaxFreqInd	Index largest frequency component
meanFreq	Frequency signal weighted average
skewnessFreq	Frequency signal Skewness
kurtosisFreq	Frequency signal Kurtosis
ampSprec	Amplitude Spectrum of the frequency signal
angle	Phase angle of the frequency signal

TABLE I: List of measures for computing feature vectors.



(a) (b) (c) (d)

Fig. 2: The ellipse criteria of segmentation.

attention maps represent the state-of-the-art tools for explanations.

B. Shallow Models

Shallow models require to properly engineer, from the raw time series, features able to adequately capture the phenomena under exam exploiting the domain knowledge. In our case, we segmented the ellipse (and then the associated time series) in different ways. Then, we extracted features from these segments based on state-of-the-art signal processing techniques [10], [31], [46]–[52] such as the mean, the median, and the signal magnitude area for both the time and frequency domains (see Table I). The Fast Fourier Transform was used to obtain frequency components for each split considered. Table I shows the list of measures applied in both time and frequency domain signals. The complete list of features is also present in the repository associated with this work[2]. The ellipse has been segmented (split) according to five different criteria (see Figure 2):

1) the ellipse is divided in segments characterised by high and low curvature (see Figure 2a);
2) the ellipse is divided in two symmetric parts according to the longest diagonal (see Figure 2b);
3) the ellipse is divided in four parts: the two more curved and the two more linear (see Figure 2c);
4) the ellipse is divided in six parts as depicted in Figure 2d;
5) all the previous split criteria (Figures 2a-2d) are considered.

On top of this feature engineering step we applied a series of state-of-the-art classification algorithms [53], [54]: RF, Support Vector Machines (with linear and Gaussian kernel), XGBoost, K-Nearest Neighbors. However, we decided to report only the results obtained using the best performing method of this family to face this problem: RF [4]. Nevertheless, the complete results are present in the repository associated with this work[2]. The RF algorithm [30] is an extension of the decision tree algorithm [55], in which decision trees are combined and each decision tree, composing the forest of n_t trees, is independently trained. The training procedure of each tree is performed as follows. First from the training dataset, a bootstrap sample is drawn as a randomised subset. Then each individual tree is grown using the randomised subset of the features (of cardinality n_f) in the creation of each node. The trees are grown until the number of elements in each leaf reaches a minimum of n_l. The accuracy of the RF converges to a maximum increasing n_t while for n_f and n_l optimal values need to be searched during the MS phase [30], [56]. In RF the full list of the hyperparameters that we tuned is reported in Table II in Section IV.

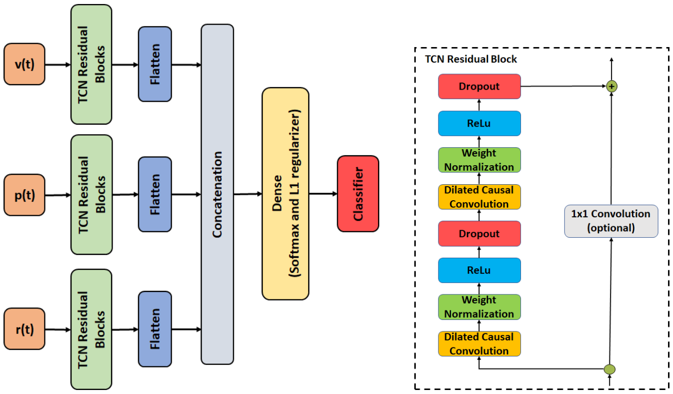
Finally, RF effectively and efficiently allows one to understand if the learning process has also a cognitive meaning, namely if it is able to capture the underline phenomena and not just capture spurious correlation [39], [40]. In particular, we decided to exploit a permutation test coupled with the Mean Decrease in Accuracy metric [57] in RF in order to identify the importance of each of the sections to provide some explanation on the learning process.

C. Deep Networks

Note that shallow models, for the scope of this work, have two main limitations. The first one is to rely on a handcrafted and experience-based feature engineering phase. The second, and most important one, is the fact that the temporal scales cannot be learned from the data, but are defined by the way we split the ellipse and the way we engineer the features. As we will describe in the rest of this section, deep models allow us to overcome both limitations.

For this purpose, we will first rely on a state-of-the-art and popular approach (i.e., classical and bidirectional LSTM [58]) to learn the features and to capture the multiple temporal scales of the problems under exam. However, we decide to report only the results obtained using the classical LSTM (since it outperformed the bidirectional LSTM) and the complete results are present in the repository associated with this work[2]. In particular, we defined a standard architecture [29], [32], [33] where the three different raw time series ($v(t)$, $r(t)$, and $p(t)$) are fed to an LSTM layer (the size of the LSTM outputs is equal to the one of the input series) which is in

[4] Note that this is something that happens in many real world problems. For example, results in Kaggle www.kaggle.com, the most popular Machine Learning competition website, show how RF and XGBoost algorithms are the top winner algorithms.



(a) High level representation of the architecture.

(b) TCN-based layer architecture details.

Fig. 3: The proposed Deep Multi Scale Models architecture based on TCN.

charge to extract the representation that is then fed directly to a dense layer which will produce the prediction. Network has been trained with the ADAM optimiser [59] empowered with cyclical learning rate [60].

This architecture has a series of hyperparameters to be tuned (see Section III-D and Table II). In particular we search the cyclical learning l_r, the dropout between the representation and the output, and the L1 regularisation in the output. Moreover, we also tuned the hyperparameters more related to the LSTM architecture such as the number of features in the hidden size and the number of recurrent layers. The complete list of the LSTM hyperparameters is reported in Table II.

Unfortunately, as we will show in the experimental results, this architecture was not able to outperform the shallow model. The reason for these results, based on the studies available in the literature, is that our shallow approach, while being naive, tries to extract features at different time scales. LSTM, instead, is only able to handle two temporal scales (a long and a short temporal scale) and this limitation will be addressed with the use of architecture based on TCN.

For this purpose we decided to rely on TCN residual blocks [34], [36] which is capable of learning different temporal scales for each raw input time series. The proposed architecture is reported in Figure 3. The peculiarities of the proposed deep multiple temporal scale architecture based on TCN are mainly three: (i) the convolutions in the architecture are causal, namely there is not information leakage from future to past, (ii) the architecture can handle different sequence lengths and map it to an output sequence of the same length as the LSTM, and (iii) is able to handle long effective history. For what concerns (i) the TCN uses causal convolutions. For what concerns (ii), it is due to the use of 1D fully-convolutional network model where each hidden layer is the same length as the input layer; zero padding of length (kernel size $-$ 1) is added to preserve the previous length. As for the (iii) we employed dilated convolution that enables a large receptive field [61] without employing too deep TCN residual blocks. The network has been trained, as for the LSTM, with the ADAM optimiser empowered with cyclical learning rate.

This architecture has a series of hyperparameters that need to be carefully tuned (see Section III-D and Table II). In particular, we explore the cyclical learning rate l_r, the dropout after the ReLU activation functions in TCN residual blocks, the L1 regularisation in dense layer, and the number of TCN residual blocks for each input data ($v(t)$, $p(t)$, and $r(t)$). Moreover, for each TCN residual block, we search the number of convolutional filters in and the kernel size of each convolutional filter. Note also that the number of residual blocks and kernel sizes are tuned in order to deliver a multiple temporal scale probing property. The complete list of the TCN hyperparameters is reported in Table II.

Finally, as for the shallow models, to understand what parts of the time series ($v(t)$, $r(t)$, and $p(t)$) mostly contribute to the decision, we decides to exploit the Gradient weighted Class Activation Mapping (Grad-CAM) [41] techniques on top of the learned model. Grad-CAM allows to easily visualise the most important part of the input time series since it extends traditional class activation maps and can be applied to a broader variety of architectures. In fact, the result of Grad-CAM is a localisation map that highlights key sections in an input for a given class, providing insights on where neural networks focus their attention.

D. Model Selection and Error Estimation

Model Selection (MS) and Error Estimation (EE) face and address the problem of tuning and assessing the performance of a learning algorithm [45]. In this work we will exploit the resampling techniques which leverage on a simple idea: $\mathcal{D}_n$ is resampled many (n_r) times, with or without replacement, and three independent datasets called learning, validation and test sets, respectively $\mathcal{L}_l^r$, $\mathcal{V}_v^r$, and $\mathcal{T}_t^r$, with $r \in \{1, \cdots, n_r\}$ are defined. Note that $\mathcal{L}_l^r \cap \mathcal{V}_v^r = \oslash$, $\mathcal{L}_l^r \cap \mathcal{T}_t^r = \oslash$, $\mathcal{V}_v^r \cap \mathcal{T}_t^r = \oslash$, and $\mathcal{L}_l^r \cup \mathcal{V}_v^r \cup \mathcal{T}_t^r = \mathcal{D}_n$ for all $r \in \{1, \cdots, n_r\}$.

Then, to select the optimal configuration of hyperparameters $\mathcal{H}$ of the algorithm $\mathscr{A}_{\mathcal{H}}$ in a set of possible ones $\mathfrak{H} = \{\mathcal{H}_1, \mathcal{H}_2, \cdots\}$, namely to perform the MS phase, the following procedure has to be applied:

$$\mathcal{H}^* : \quad \arg\min_{\mathcal{H} \in \mathfrak{H}} \sum_{r=1}^{n_r} M(\mathscr{A}_{\mathcal{H}}(\mathcal{L}_l^r), \mathcal{V}_v^r), \tag{1}$$

where $\mathscr{A}_{\mathcal{H}}(\mathcal{L}_l^r)$ is a model learned by $\mathscr{A}$ with the hyperparameters $\mathcal{H}$ based on the the data in $\mathcal{L}_l^r$ and where $M(f, \mathcal{V}_v^r)$ is the desired metric computed for f on $\mathcal{V}_v^r$. Since the data in $\mathcal{L}_l^r$ are independent from the ones in $\mathcal{V}_v^r$, the intuition is that $\mathcal{H}^*$ should be the configuration of hyperparameters which allows to achieve optimal performance, according to the desired metric, on a set of data that is independent, namely previously unseen, with respect to the training set.

Then, in order to evaluate the performance of the optimal model, namely the model learned with the optimal hyperparameters based on the available data, which is $f^*_{\mathscr{A}} = \mathscr{A}_{\mathcal{H}^*}(\mathcal{D}_n)$ or, in other words, to perform the EE phase, the following procedure has to be applied:

$$M(f^*_{\mathscr{A}}) = \frac{1}{n_r} \sum_{r=1}^{n_r} M(\mathscr{A}_{\mathcal{H}^*}(\mathcal{L}_l^r \cup \mathcal{V}_v^r), \mathcal{T}_t^r). \tag{2}$$

Algorithm	Hyperparameters
RF	$n_f : \{d^{1/3}, d^{1/2}, d^{3/4}\}$ $n_l : \{1\}$ $n_t : \{1000\}$
LSTM	l_r: $\{0.0001, 0.0005, 0.001\}$ dropout: $\{0.1, 0.15, \dots, 0.5\}$ L1 regularization: $\{0.0001, 0.0005, 0.001, 0.005\}$ hidden dimensions: $\{25, 50, 75, 111\}$ hidden layers: $\{1, 2, 3, 4\}$
TCN	l_r: $\{0.0001, 0.0005, 0.001\}$ dropout: $\{0.1, 0.15, \dots, 0.5\}$ L1 regularization: $\{0.0001, 0.0005, 0.001, 0.005\}$ hidden layers $\{1, 2, 3\}$ convolutional filters: $\{32, 64, 128\}$ kernel dimensions: $\{3, 5, 7, 9, 11\}$

TABLE II: Hyperparameters for all algorithms tested in our analysis.

Since the data in $\mathcal{L}_l^r \cup \mathcal{V}_v^r$ are independent from the ones in $\mathcal{T}_t^r$, $M(\mathscr{A}_{\mathcal{H}^*}(\mathcal{L}_l^r \cup \mathcal{V}_v^r, \mathcal{T}_t^r)$ will be an unbiased estimator of the true performance of the final model [45].

In this paper the complete k-fold cross validation is exploited [45], [62] since, together with bootstrap, represent a state-of-the-art approach to the problem of MS and EE [62].

E. Scenarios

In our experiment, we studied two scenarios in order to understand the extrapolation capability of the different model described in Sections III-B and III-C:

- Leave One Hand of one subject Out (LOHO): in this scenario the models have been trained with all the subjects data except the ones related to one hand of one subject which has been kept apart for testing purposes;
- Leave One Speed of one subject Out (LOSO): in this scenario the models have been trained with all the subjects data except the ones related to one speed of one subject which has been kept apart for testing purposes;

Therefore, the two scenarios just differ in the definition of the three sets $\mathcal{L}_l^r$, $\mathcal{V}_v^r$, and $\mathcal{T}_t^r$, which are the subset of data employed for building, tuning, and testing the models. For instance, in the LOHO scenario $\mathcal{L}_l^r$, $\mathcal{V}_v^r$, and $\mathcal{T}_t^r$ have been created by randomly selecting data from one hand of one subject to be inserted in $\mathcal{T}_t^r$, from another hand of a different subject to be inserted in $\mathcal{V}_v^r$, and from the remaining ones to be inserted into $\mathcal{L}_l^r$.

IV. Experimental Results

In this section, we will report the results of applying the methodology presented in Section III over the data described in Section II.

The full list of the hyperparameters tested is reported in Table II.

Table III reports the percentage of accuracy, in the LOHO and LOSO scenarios respectively, when exploiting RF (with the different criteria of segmentation of the ellipses described in Section III-B), LSTM, and TCN (see Section III-C) for each of the 14 subjects together with the average across the subjects.

Table III allows to observe that:

Alg. / Subj.	RF (a)	RF (b)	RF (c)	RF (d)	RF (e)	LSTM	TCN
1	98.0±0.1	98.4±0.2	99.7±0.2	100.0±0.1	100.0±0.1	90.5±1.5	97.8±0.7
2	99.1±0.1	99.4±0.2	99.6±0.1	99.9±0.1	99.9±0.1	83.9±2.1	99.1±0.2
3	96.6±0.5	97.6±0.4	98.2±0.3	97.9±0.3	97.1±0.5	92.8±2.2	98.2±1.0
4	68.1±1.5	71.3±1.5	70.7±2.3	69.4±2.9	71.0±1.9	85.8±2.2	86.9±1.7
5	99.8±0.1	99.8±0.1	99.8±0.1	100.0±0.1	100.0±0.1	92.2±3.1	98.9±0.2
6	75.7±2.3	91.5±0.9	81.2±1.6	75.3±1.4	92.5±1.0	64.7±3.5	90.4±1.0
7	99.0±0.1	97.8±0.8	99.9±0.1	100.0±0.1	86.0±0.1	91.6±2.2	98.7±0.3
8	100.0±0.1	100.0±0.1	100.0±0.1	100.0±0.1	100.0±0.1	85.1±3.5	99.2±0.4
9	98.3±0.3	98.6±0.5	99.8±0.2	100.0±0.1	99.9±0.1	90.0±1.3	97.5±0.7
10	98.1±0.3	98.5±0.6	99.8±0.1	100.0±0.1	100.0±0.1	87.6±1.4	98.6±0.7
11	98.7±0.2	98.7±0.3	99.7±0.2	99.9±0.1	99.8±0.1	86.0±2.6	99.0±0.1
12	97.8±0.1	98.2±0.3	99.5±0.7	99.8±0.5	99.6±0.4	90.9±1.7	97.6±0.4
13	98.9±0.2	98.9±0.1	99.4±0.1	99.7±0.1	99.7±0.1	92.4±1.8	98.4±0.7
14	98.4±0.2	98.3±0.4	99.9±0.1	100.0±0.1	99.9±0.1	93.5±1.8	99.3±0.3
Avg.	**94.8±0.4**	**96.2±0.5**	**96.2±0.4**	**95.9±0.4**	**96.1±0.3**	**87.6±2.2**	**97.1±0.6**

(a) LOHO

Alg. / Subj.	RF (a)	RF (b)	RF (c)	RF (d)	RF (e)	LSTM	TCN
1	98.7±0.2	99.2±0.2	100.0±0.1	100.0±0.1	100.0±0.1	97.6±0.7	99.7±0.2
2	99.1±0.1	99.4±0.2	99.5±0.1	100.0±0.1	99.7±0.1	92.4±2.1	99.3±0.3
3	97.7±0.2	98.9±0.2	98.2±0.1	98.3±0.1	96.8±0.2	97.8±1.0	98.2±0.5
4	81.1±0.8	96.8±0.5	90.9±0.8	92.6±0.6	91.3±0.3	97.0±0.6	96.3±0.9
5	99.7±0.1	99.7±0.1	99.9±0.1	100.0±0.1	100.0±0.1	98.1±0.8	99.6±0.4
6	85.9±1.0	93.8±1.3	87.7±0.6	84.7±0.8	89.7±0.1	96.7±1.4	96.7±0.7
7	99.5±0.1	99.6±0.1	100.0±0.1	100.0±0.1	99.5±0.1	95.9±0.8	99.2±0.4
8	100.0±0.1	100.0±0.1	100.0±0.1	100.0±0.1	100.0±0.1	93.8±2.9	99.5±0.5
9	98.4±0.1	99.2±0.2	99.8±0.1	99.9±0.1	100.0±0.1	97.9±0.4	99.6±0.5
10	98.9±0.2	99.3±0.2	100.0±0.1	100.0±0.1	100.0±0.1	95.2±0.8	99.4±0.2
11	98.9±0.1	99.5±0.1	99.8±0.1	100.0±0.1	99.8±0.1	96.8±0.6	99.7±0.3
12	97.9±0.1	98.3±0.2	99.7±0.3	99.7±0.3	98.2±0.1	97.8±0.7	99.7±0.2
13	99.0±0.1	99.0±0.1	99.6±0.1	99.8±0.1	97.8±0.1	98.6±0.5	99.9±0.1
14	98.2±0.1	99.2±0.3	100.0±0.1	100.0±0.1	100.0±0.1	99.5±0.3	100.0±0.1
Avg.	**96.7±0.2**	**98.7±0.3**	**98.2±0.2**	**98.2±0.2**	**98.1±0.1**	**96.8±1.0**	**99.1±0.4**

(b) LOSO

TABLE III: ACC when exploiting RF (with the different criteria of segmentation of the ellipses described in Section III-B), LSTM, and TCN (see Section III-C) for each of the 14 subjects together with the average across the subjects.

- As one might expect, performances on LOSO are generally higher than the ones on LOHO for all subjects and algorithms. This is the natural consequence of the fact that in LOHO we are asking a more complex extrapolation capability to the algorithms;
- TCN consistently outperforms RF and LSTM in all scenarios while demonstrating also consistent performance across the subjects;
- RF is quite competitive and outperforms, for some subjects, also TCN. Nevertheless, for some subjects, performance are quite poor;
- RF in case (a) and (b) performs quite well. These results indicate that segmenting too little or too much the ellipse is not a good solution while putting all the possible segmentation, as in case (e), does not guarantee optimal performance. In fact, these segmentations designed to capture multiple temporal scales are, by construction, fixed and not customised for the specific problem. The

Met. \ Alg.	RF (a)	RF (b)	RF (c)	RF (d)	RF (e)	LSTM	TCN
ACC	94.8±0.4	96.2±0.5	96.2±0.4	95.9±0.4	96.1±0.3	87.6±2.2	97.1±0.6
REC	94.8±0.4	96.2±0.5	96.2±0.4	95.9±0.4	96.1±0.3	87.6±2.2	97.1±0.6
PRE	94.8±0.3	96.2±0.5	96.2±0.4	95.9±0.4	96.1±0.4	87.6±2.5	97.1±0.4
ROC-AUC	0.9±0.1	0.9±0.1	0.9±0.1	0.9±0.1	0.9±0.1	0.9±0.1	0.9±0.1

(a) LOHO

Met. \ Alg.	RF (a)	RF (b)	RF (c)	RF (d)	RF (e)	LSTM	TCN
ACC	96.7±0.2	98.7±0.3	98.2±0.2	98.2±0.2	98.1±0.1	96.8±1.0	99.1±0.4
REC	96.7±0.2	98.7±0.3	98.2±0.2	98.2±0.2	98.1±0.1	96.8±1.0	99.1±0.4
PRE	96.7±0.2	98.7±0.3	98.2±0.2	98.3±0.2	98.1±0.2	96.8±1.2	99.2±0.1
ROC-AUC	0.9±0.1	0.9±0.1	0.9±0.1	0.9±0.1	0.9±0.1	0.9±0.1	0.9±0.1

(b) LOSO

TABLE IV: ACC, REC, PRE, and ROC-AUC, averaged over the 14 subjects when exploiting RF (with the different criteria of segmentation of the ellipses described in Section III-B), LSTM, and TCN (see Section III-C).

TCN-based architecture, instead, actually learns the correct temporal scale to focus on;

- LSTM, as expected, is the algorithm which demonstrates the lowest performance. This is due to the fact that its ability to capture different temporal scales is too limited.

For completeness, we report ACC, REC, PRE and ROC-AUC in Table IV averaged over the 14 subjects.

As mentioned in Section III-B, we did not report in the paper the results with other shallow algorithms (e.g., Support Vector Machines) or the deep one (e.g., bidirectional LSTM) since they under-perform RF, LSTM, and the TCN-based architecture. We reported instead LSTM to show that naive deep architectures do not outperform classical methods as RF. Nevertheless, the complete set of results can be found in our publicly available repository[2].

As described in Section III, in order to better understand how and what the different RF and TCN models actually learned from the data, Table V reports the sections ranking [5] performed with RF in the different sectioning scenarios (see Section III-B) and Figure 4 reports the attention maps of TCN (see Section III-C), averaged across subjects, for $p(t)$, $v(t)$, and $r(t)$ for both LOHO and LOSO scenarios.

Table V and Figure 4 allow to observe that:

- As one might expect, the most important sections in the two scenarios (LOHO and LOSO) do not result to be exactly the same since they try to extrapolate with respect to different information (hand and speed). When using shallow models (i.e., RF) for sectioning (a), (b), and (c) sections maintain the same importance in both LOHO and LOSO scenarios while for sectioning (d) and (e) the ranking is quite different. When using deep models (i.e., TCN), instead, only for $v(t)$ the attention map remain similar for both LOHO and LOSO scenarios;

[5] The letters indicate the sectioning and the numbers indicate the specific section, see Figure 2, so note that c.1 is the same as d.1 and c.3 is the same as d.4.

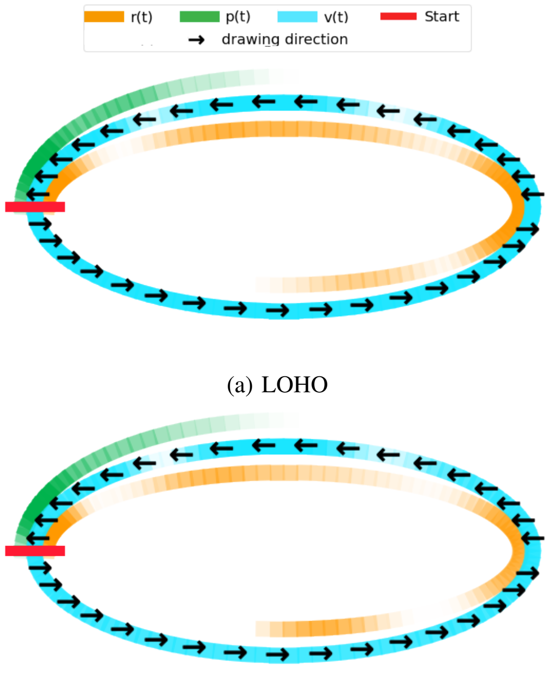
Sectioning	Rank 1	2	3	4	5	6	7	8	9	10	11	12
(a)	a.2	a.1										
(b)	b.1	b.2										
(c)	c.4	c.2	c.1	c.3								
(d)	d.3	d.2	d.4	d.1	d.5	d.6						
(e)	d.3	d.2	c.1 (d.1)	c.3 (d.4)	b.1	c.2	d.6	d.5	a.2	b.2	a.1	c.4

(a) LOHO

Sectioning	Rank 1	2	3	4	5	6	7	8	9	10	11	12
(a)	a.2	a.1										
(b)	b.1	b.2										
(c)	c.4	c.2	c.1	c.3								
(d)	d.2	d.4	d.3	d.1	d.5	d.6						
(e)	c.1 (d.1)	d.3	c.2	a.1	b.2	b.1	d.6	a.2	d.5	c.3 (d.4)	c.4	d.2

(b) LOSO

TABLE V: Sections ranking[5] performed with RF in the different sectioning scenarios for both LOHO and LOSO scenarios.



(a) LOHO

(b) LOSO

Fig. 4: Attention maps of TCN, averaged across subjects, for $v(t)$, $r(t)$, and $p(t)$ for both LOHO and LOSO scenarios. The more intense the colour, the more important is the particular part of the input time series.

- For both LOHO and LOSO scenarios, shallow models identify as the most informative sections those who are closer to the initial part of the drawing in all the analysed sectioning criteria. On the other hand, deep models generally find the final parts of the drawing as most informative. This shows how different the perception of the two models is. The shallow ones focus on the "preparation" of the movement, while the deep ones focus more on the "completion" of the movement. The deep model, in this case, perceives the movement in a way which seems more similar to a human: human beings tend to become more confident in labelling a movement when it tends to be completed;
- shallow models primarily focus on more "linear" sections with respect to the more "curved ones". The opposite happens for deep models. Also in this case, deep model perception is more similar to human one: human tends to distinguish movements based on the most complex parts;
- Finally note that shallow models tend to focus on sections based on the particular choice of the sectioning criteria and are not able to perceive and define their one way of understanding the movement. Deep models, instead, by construction are able to do so by defining the attention maps based on the particular problem and defining implicitly their own sectioning criteria then being able to perceive the different time scales of the movement.

V. Conclusion

In this paper we investigated how shallow and deep data-driven models can be of support in understanding the human movement, and in particular its multiple temporal scale nature. We showed how shallow data-driven models, which achieve reasonably good recognition performance, require a usually complex phase of handcrafting of the features, based on domain-specific knowledge, thus limiting the ability to extract all the possible information from the data. For this reason, we propose a new deep multi-scale data-driven model based on temporal convolutional networks able to automatically learn features from the data at different temporal scales and outperforms state-of-the-art shallow models in terms of recognition performance. We tested the effectiveness of our approach in a custom experiment of movement recognition, namely detecting the person who draws an ellipse on a graphics tablet based on the velocity, pressure, and curvature of the drawing movement. Exploiting the intrinsic hierarchy in the dataset, we considered two different extrapolation scenarios, namely on hand and on speeds, to understand the potentiality of the proposed architecture. Results, both in terms of performance and interpretability of the result, support the need and the usefulness of studying human movement at different temporal scales by means of multi temporal scale data-driven models. In particular, We observed how the differences between the proposed model and a traditional one are not so evident in terms of recognition performance but they are when it comes to understanding what the different models actually learned from data. Traditional models tend to perform very well for certain subjects and poorly perform on other subjects and the information extracted from the data is usually different from the human intuition. The proposed architecture, instead, tends to perform well consistently across subjects and to extract information more in agreement with human intuition. In addition, all the data and code related to this study have been made freely available for the community[2]. Nevertheless, this is just a first step forward in understanding the benefits of a machine learning model based on human perception. In fact, we plan to enlarge the experimental setup and to test our approach with a larger number of subjects to further support the results presented in this work.

Acknowledgements

This research has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824160 (FET PROACTIVE EnTimeMent Project).

References

[1] J. P. Piek, *Motor behavior and human skill: a multidisciplinary approach*. Human Kinetics, 1998.

[2] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 15–33, 2012.

[3] S. Piana, A. Staglianò, F. Odone, and A. Camurri, "Adaptive body gesture representation for automatic emotion recognition," *ACM Transactions on Interactive Intelligent Systems*, vol. 6, no. 1, pp. 1–31, 2016.

[4] R. Szeliski, *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[5] Q. Lei, J. X. Du, H. B. Zhang, S. Ye, and D. S. Chen, "A survey of vision-based human action evaluation methods," *Sensors*, vol. 19, no. 19, p. 4129, 2019.

[6] N. De Giorgis, E. Puppo, P. Alborno, and A. Camurri, "Evaluating movement quality through intrapersonal synchronization," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 4, pp. 304–313, 2019.

[7] A. O. Holcombe, "Seeing slow and seeing fast: two limits on perception," *Trends in cognitive sciences*, vol. 13, no. 5, pp. 216–221, 2009.

[8] S. Sepp, S. J. Howard, S. Tindall-Ford, S. Agostinho, and F. Paas, "Cognitive load theory and human movement: Towards an integrated model of working memory," *Educational Psychology Review*, vol. 31, pp. 1–25, 2019.

[9] E. Halilaj, A. Rajagopal, M. Fiterau, J. L. Hicks, T. J. Hastie, and S. L. Delp, "Machine learning in human movement biomechanics: Best practices, common pitfalls, and new opportunities," *Journal of biomechanics*, vol. 81, pp. 1–11, 2018.

[10] V. D'Amato, E. Volta, L. Oneto, G. Volpe, A. Camurri, and D. Anguita, "Understanding violin players' skill level based on motion capture: a data-driven perspective," *Cognitive Computation*, vol. 12, no. 6, pp. 1356–1369, 2020.

[11] B. Abernethy, *Biophysical foundations of human movement*. Human Kinetics, 2013.

[12] R. M. Enoka, *Neuromechanics of human movement*. Human kinetics, 2008.

[13] M. Wijnants, R. Cox, F. Hasselman, A. Bosman, and G. Van Orden, "A trade-off study revealing nested timescales of constraint," *Frontiers in physiology*, vol. 3, p. 116, 2012.

[14] J. Butepage, M. J. Black, D. Kragic, and H. Kjellstrom, "Deep representation learning for human motion prediction and classification," in *IEEE conference on computer vision and pattern recognition*, 2017.

[15] F. Zhou, F. De la Torre, and J. K. Hodgins, "Hierarchical aligned cluster analysis for temporal clustering of human motion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 582–596, 2012.

[16] D. Aur, "From neuroelectrodynamics to thinking machines," *Cognitive Computation*, vol. 4, pp. 4–12, 2012.

[17] R. L. Goldstone, J. R. de Leeuw, and D. H. Landy, "Fitting perception in and to cognition," *Cognition*, vol. 135, pp. 24–29, 2015.

[18] G. Johansson, "Visual perception of biological motion and a model for its analysis," *Perception & psychophysics*, vol. 14, no. 2, pp. 201–211, 1973.

[19] F. E. Pollick, H. M. Paterson, A. Bruderlin, and A. J. Sanford, "Perceiving affect from arm movement," *Cognition*, vol. 82, no. 2, pp. B51–B61, 2001.

[20] H. K. M. Meeren, N. Hadjikhani, S. P. Ahlfors, M. S. Hämäläinen, and B. De Gelder, "Early preferential responses to fear stimuli in human right dorsal visual stream-a meg study," *Scientific reports*, vol. 6, p. 24831, 2016.

[21] E. Ihlen and B. Vereijken, "Interaction-dominant dynamics in human cognition: Beyond $1/f\alpha$ fluctuation." *Journal of Experimental Psychology: General*, vol. 139, no. 3, p. 436, 2010.

[22] A. Camurri, G. Volpe, S. Piana, M. Mancini, R. Niewiadomski, N. Ferrari, and C. Canepa, "The dancer in the eye: towards a multi-layered computational framework of qualities in movement," in *Proceedings of the 3rd International Symposium on Movement and Computing*, 2016.

[23] J. Newlove and J. Dalby, *Laban for all*. Routledge, 2019.

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.

[25] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, pp. 834–848, 2017.

[26] X. Dai, B. Singh, J. Ng, and L. Davis, "Tan: Temporal aggregation network for dense multi-label action recognition," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.

[27] C. Beyan, S. Karumuri, G. Volpe, A. Camurri, and R. Niewiadomski, "Modeling multiple temporal scales of full-body movements for emotion classification," *IEEE Transactions on Affective Computing*, 2021.

[28] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory To Algorithms*. Cambridge University Press, 2014.

[29] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[30] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[31] A. Roy, B. Banerjee, A. Hussain, and S. Poria, "Discriminative dictionary design for action classification in still images and videos," *Cognitive Computation*, vol. 13, pp. 698–708, 2021.

[32] D. Liciotti, M. Bernardini, L. Romeo, and E. Frontoni, "A sequential deep learning application for recognising human activities in smart homes," *Neurocomputing*, vol. 396, pp. 501–513, 2020.

[33] D. Jirak, S. Tietz, H. Ali, and S. Wermter, "Echo state networks and long short-term memory for continuous gesture recognition: A comparative study," *Cognitive Computation*, pp. 1–13, 2020.

[34] S. M. Lee, S. M. Yoon, and H. Cho, "Human activity recognition from accelerometer data using convolutional neural network," in *IEEE international conference on big data and smart computing*, 2017.

[35] G. Buzsáki and A. Draguhn, "Neuronal oscillations in cortical networks," *science*, vol. 304, no. 5679, pp. 1926–1929, 2004.

[36] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[37] G. Wen, Z. Hou, H. Li, D. Li, L. Jiang, and E. Xun, "Ensemble of deep neural networks with probability-based fusion for facial expression recognition," *Cognitive Computation*, vol. 9, no. 5, pp. 597–610, 2017.

[38] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[39] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[40] C. S. Calude and G. Longo, "The deluge of spurious correlations in big data," *Foundations of science*, vol. 22, no. 3, pp. 595–612, 2017.

[41] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *IEEE international conference on computer vision*, 2017.

[43] P. Viviani and C. Terzuolo, "Trajectory determines movement dynamics," *Neuroscience*, vol. 7, no. 2, pp. 431–437, 1982.

[42] L. Scocchia, N. Bolognini, S. Convento, and N. Stucchi, "Cathodal transcranial direct current stimulation can stabilize perception of movement: evidence from the two-thirds power law illusion," *Neuroscience letters*, vol. 609, pp. 87–91, 2015.

[44] M. Badarna, I. Shimshoni, G. Luria, and S. Rosenblum, "The importance of pen motion pattern groups for semi-automatic classification of handwriting into mental workload classes," *Cognitive Computation*, vol. 10, no. 2, pp. 215–227, 2018.

[45] L. Oneto, *Model Selection and Error Estimation in a Nutshell*. Springer, 2019.

[46] N. D. Cilia, C. De Stefano, F. Fontanella, and A. Scotto, "A ranking-based feature selection approach for handwritten character recognition," *Pattern Recognition Letters*, vol. 121, pp. 77–86, 2019.

[47] P. Drotár, J. Mekyska, I. Rektorová, L. Masarová, Z. Smékal, and M. Faundez-Zanuy, "Evaluation of handwriting kinematics and pressure for differential diagnosis of parkinson's disease," *Artificial intelligence in Medicine*, vol. 67, pp. 39–46, 2016.

[48] A. Sama, D. E. Pardo-Ayala, J. Cabestany, and A. Rodríguez-Molinero, "Time series analysis of inertial-body signals for the extraction of dynamic properties from human gait," in *International Joint Conference on Neural Networks*, 2010.

[49] N. Wang, E. Ambikairajah, N. H. Lovell, and B. G. Celler, "Accelerometry based classification of walking patterns using time-frequency analysis," in *IEEE Engineering in Medicine and Biology Society*, 2007.

[50] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in *International conference on pervasive computing*, 2004.

[51] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine," in *International workshop on ambient assisted living*, 2012.

[52] K. Rapantzikos, Y. Avrithis, and S. Kollias, "Spatiotemporal features for action recognition and salient event detection," *Cognitive Computation*, vol. 3, pp. 167–184, 2011.

[53] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3133–3181, 2014.

[54] M. Wainberg, B. Alipanahi, and B. J. Frey, "Are random forests truly the best classifiers?" *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 3837–3841, 2016.

[55] L. Rokach and O. Z. Maimon, *Data Mining with Decision Trees: Theory and Applications*. World Scientific, 2008, vol. 69.

[56] I. Orlandi, L. Oneto, and D. Anguita, "Random forests model selection," in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2016.

[57] R. Genuer, J. M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2225–2236, 2010.

[58] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks," in *Proceedings of the AAAI conference on artificial intelligence*, 2016.

[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[60] L. N. Smith, "Cyclical learning rates for training neural networks," in *IEEE winter conference on applications of computer vision*, 2017.

[61] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[62] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *International Joint Conference on Artficial Intelligence*, 1995.