AN UNSUPERVISED APPROACH TO THE SEMANTIC DESCRIPTION OF THE SOUND QUALITY OF VIOLINS

M. Buccoli, M. Zanoni, F. Setragno, F. Antonacci, A. Sarti

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano
Piazza Leonardo da Vinci 32 - 20133 Milano, Italy

{michele.buccoli, massimiliano.zanoni, francesco.setragno,
fabio.antonacci, augusto.sarti}@polimi.it

ABSTRACT

In this study we propose a set of semantic musical descriptors that can be used for describing the timbre of violins. The proposed semantic model follows a dimensional approach, which allows us to express the degree of intensity of each descriptor. A set of recordings of a number of violins (among them, Stradivari, Amati and Guarnieri instruments) were annotated with the descriptors through questionnaires. The recordings are processed with deep learning techniques, to learn salient features from the audio signal in an unsupervised fashion. In this study we propose an automatic annotation procedure based on a set of regression functions that model each semantic descriptor using the learned set of features.

Index Terms— High-level music descriptor, violin, timbre, sound quality

1. INTRODUCTION

The study of the timbral qualities of violins has been the subject of intense scientific investigation [1] for decades. However, the physical phenomena that are involved in the characterization of their timbral quality are still far from being fully understood [2].

Classical approaches to the study of sound proprieties of musical instruments consists of extracting and analyzing a large set of acoustic cues (*Low-Level Features - LLF*) [3]. Concerning the characterization of the sound quality violins, in [4, 5] the authors use a set of MPEG spectral and harmonic descriptors, whereas in [6], the author uses long-term cepstral coefficients. Such descriptors, however, are characterized by a low level of abstraction. Musicians and instrument makers, in fact, tend to describe the sound quality of their instruments using terms coming from natural language (e.g. *warm*, *bright*, ...), and are therefore semantically rich [7]. This is why in our work we focused on *Semantic Timbral Descriptors*, or High-Level Features (HLF).

This research activity has been partially funded by the Cultural District of the province of Cremona, Italy, a *Fondazione CARIPLO* project, and by the Arvedi-Buschini Foundation

Though semantic timbral descriptions are inherently subjective, there exists a strong connection between sound description, sound perception and physics. Our brain, in fact, processes stimuli from the auditory system in order to formulate a proper description. Understanding which aspects of the sound influence our perception [8], however, is not an easy task. For this reason, although some work has been done in this direction [8–10], this connection is still not fully understood. In the literature this is known as the *semantic gap between Low-Level and High-Level Features*. In the past few year, the Music Information Retrieval (MIR) community has focused a great deal on techniques to fill this gap, particularly for the automatic semantic annotation of musical content [11]. In the area of musical acoustics, some studies have already appeared in the literature, which focus on the semantic description of the violin timbre [12–14]. In [2], the correlation between LLF and HLF was studied using a set of correlation indexes. In that work, machine learning techniques were employed for modeling Semantic Descriptors for automatic annotation and retrieval. In particular, generative solutions based on regression analysis were employed, which is an approach that was recently applied to Music Emotion Recognition [15–17] with remarkable results.

In order to build the model for semantic descriptors we need to collect low-level and high-level representations of a large set of instruments. In order to do so, we recorded numerous violins, some of which were historical instruments from the collection of the Museo del Violino in Cremona, Italy (made by Antonio Stradivari, Giuseppe Guarnieri "del Gesù" and Nicolò Amati). As far as the low-level representation is concerned, one typical approach consists of manually selecting the most relevant (discriminant) features for the task at hand [3, 8]. In this study, however, we followed a different approach, in which the acoustic cues are "learnt" directly from the available data using *deep learning techniques* [18].

Deep learning techniques are based on an inherently layered representation of the information, which enables the inferral of features that describe the input data (*learned* features) at various levels of abstraction, in an unsupervised fash-

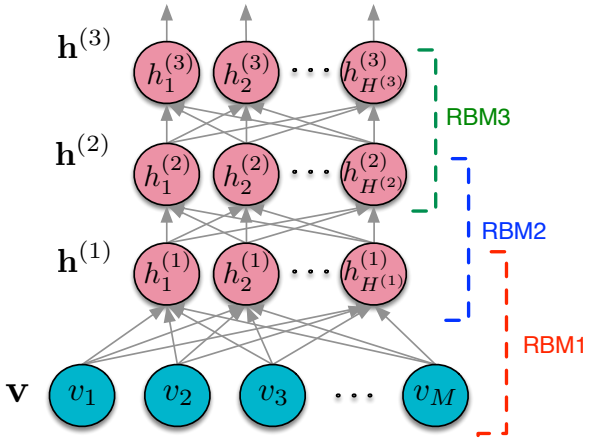

Fig. 1. Representation of a DBN, which is composed by several stacked RBM.

ion. This is an approach whose effectiveness has been proven in various tasks such as music emotion recognition [17] and categorical audio classification [19]. The set of *learned* features that are used for the modeling of HLFs follows a classical training-based approach. Machine learning regressions allow us to adopt a *dimensional* representation for the semantic descriptors, which express the degree of intensity of each descriptor [8, 15, 20].

2. UNSUPERVISED FEATURE LEARNING

Deep learning [18] refers to a set of machine-learning techniques that are based on multi-layer architectures to process and represent input data with multiple levels of abstraction. It aims at reproducing the way the human brain processes information in a hierarchical fashion to address decisional problems decomposing them into simpler sub-problems.

In this study we use the Deep Belief Networks (DBN) [18], which are composed by stacking layers of Restricted Boltzman Machines (RBM) (Fig. 1), in order to provide a hierarchical transformation of audio input data learned in an unsupervised manner.

2.1. Restricted Boltzman Machine

An RBM is composed by two layers of neurons: an input layer v and a hidden layer h. The neurons are fully connected between different layers, whereas neurons of the same layer are not connected.

The hidden layer h is trained in order to reconstruct the input layer v by minimizing an energy-based function. More formally, given a vector of input data $\mathbf{v} \in \mathbb{R}^{M\times 1}$, a vector $\mathbf{h} \in \mathbb{R}^{H\times 1}$, we define the energy of a RBM configuration as

$$E(\mathbf{v},\mathbf{h}) = -\mathbf{b}^\top\mathbf{v} - \mathbf{c}^\top\mathbf{h} - \mathbf{h}^\top\mathbf{W}\mathbf{v}, \tag{1}$$

where $W \in \mathbb{R}^{H\times M}$ is the matrix of weights that connects each input neuron to each hidden neuron, $\mathbf{c} \in \mathbb{R}^{H\times 1}$ is a bias term for the hidden neurons and $\mathbf{b} \in \mathbb{R}^{M\times 1}$ is a bias term for the visible neurons.

We can estimate the parameters of the RBM by minimizing the *free energy function*

$$F(\mathbf{v}) = -\log \sum_{\mathbf{h}\in\mathcal{H}_\mathbf{v}} e^{-E(\mathbf{v},\mathbf{h})}, \tag{2}$$

where $\mathcal{H}_\mathbf{v}$ is the set of hidden vector $\mathbf{h}$ that can be obtained from the visible vector $\mathbf{v}$. The optimal parameters are then estimate by minimizing the free energy function over a set $\mathcal{V}$ of training input vectors $\mathbf{v}$:

$$\{\hat{\mathbf{W}},\hat{\mathbf{b}},\hat{\mathbf{c}}\} = \underset{\mathbf{W},\mathbf{b},\mathbf{c}}{\operatorname{argmin}} \prod_{\mathbf{v}\in\mathcal{V}} F(\mathbf{v}). \tag{3}$$

This minimization problem is typically solved using iterative approaches [18].

The estimated parameters $\hat{\mathbf{W}}$ and $\hat{\mathbf{c}}$ are finally used to compute a representation $\mathbf{h}$ of new data as

$$\mathbf{h} = \mathcal{T}\left(\hat{\mathbf{W}}\mathbf{v} + \hat{\mathbf{c}}\right), \tag{4}$$

where $\mathcal{T}$ is a non-linear operation applied element-wise.

2.2. Deep Belief Network

As it is shown in Fig. 1, in the DBN, each RBM receives as input the hidden vector of the previous one. The RBMs are sequentially trained from the bottom layer to the top layer.

Hence, given a generic input vector $\mathbf{v}$ and a DBN composed by K layers of RBMs, the parameters $\hat{\mathbf{W}}^{(k)}$ and $\hat{\mathbf{c}}^{(k)}$ are estimated for each layer $k = 1, ..., K$ and the k-th feature vector $\mathbf{h}^{(k)} \in \mathbb{R}^{H^{(k)}}$ is extracted as:

$$\mathbf{h}^{(k)} = \mathcal{T}\left(\hat{\mathbf{W}}^{(k)}\mathbf{h}^{(k-1)} + \hat{\mathbf{c}}^{(k)}\right), \tag{5}$$

where $\mathbf{h}^{(\mathbf{0})} = \mathbf{v}$ and in this study $\mathcal{T}$ is the sigmoid function [18].

3. TIMBRE DESCRIPTION MODELING

In order to characterize the sound quality of violins, in this study we selected a set of $N_D = 6$ bipolar semantic descriptors from [2]. In [2] the authors investigated the most used terms by means of survey to a large set of violin makers. The selected descriptors are listed in Table 1. Each descriptor represents an aspect of the sound quality and it is modeled through a mono-dimensional spaces $\mathcal{D}_i \subseteq \mathcal{R}$ with $i = 1, ..., N_D$ and a pair of terms $t_i^{(l)}$ and $t_i^{(h)}$: a term and its opposite. The terms represent the two extremes of the space.

Using this formalism, each instrument s is described by the compact representation $\mathbf{y}_s = [\omega_{s,1}, \omega_{s,2}, ..., \omega_{s,N_D}]$,

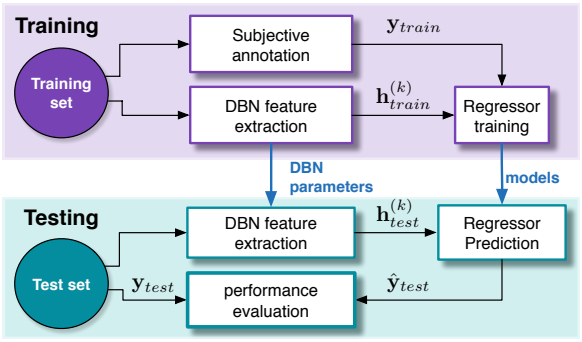

Fig. 2. Overall scheme of the semantic descriptor modeling.

where $\omega_i \in [0, 1]$ is the degree of the descriptiveness of the i-th property. A low value of ω_i, down to 0, means that the violin can be properly described by $t_i^{(l)}$, whereas a higher value, up to 1, represent that it can be described by $t_i^{(h)}$.

Each semantic descriptor is modeled following a classic schema of a training-based technique. Figure 2 shows the workflow of the method. The low-level characterization of each recording is provided through the extraction of the set of learned LLFs. Semantic descriptors are modeled through a set of generative models (regressors) that are trained on the high dimensional learned feature space computed on a training dataset of recordings.

The algorithm is validated by splitting the dataset of recordings of the violins $\mathcal{S}$ into the training set $\mathcal{S}_{train}$ and the test set $\mathcal{S}_{test}$.

3.1. Feature Extraction

Each audio recording $\mathbf{x}_s$ with $s \in \mathcal{S}$ is divided into a set of F overlapping frames $\mathbf{x}_{s,f}$ of fixed length in time, with $f = 1, ..., F$. The log-magnitude representations of the frequency spectrum of the frames are provided:

$$\mathbf{v}_{s,f} = \log_{10}(|\mathcal{F}(\mathbf{x}_{s,f})|), \tag{6}$$

where $\mathcal{F}$ is the Fourier transform. In the training phase, the set of $\mathbf{v}_{z,f}$ for $z \in \mathcal{S}_{train}$ is used to estimate the parameters of the network $\{\hat{\mathbf{W}}^{(k)}, \hat{\mathbf{c}}^{(k)}\}$ (eq. 3). The estimated parameters are used to provide the learned representation $\mathbf{h}^{(k)}_{s,f} \in \mathbb{R}^{H^{(k)}}$ for either training and test sets, at each k-th abstraction layer.

In order to obtain a more compact representation that is more robust to rapid variation, the feature vectors are averaged over sliding overlapping 5s long segments $\mathbf{h}^{(k)}_{s,a}$ where $a = 1, ..., A$ is the index of the sliding segment.

3.2. Content-based music description

Regression analysis can be used to predict a real value from a set of observed variable by projecting a multidimensional feature space into a novel continuous space with a limited number of dimensions [15]. In our case, for each semantic descriptor, the LLF space is mapped into a novel conceptual one-dimensional space of real values (HLF).

$t_i^{(l)}$	$t_i^{(h)}$	$t_i^{(l)}$	$t_i^{(h)}$	$t_i^{(l)}$	$t_i^{(h)}$
Dark	Bright	Not Deep	Deep	Not Full	Full
Hard	Soft	Not Warm	Warm	Harsh	Sweet

Table 1. Set of bipolar descriptors for violin timbre description.

Formally, given $(\mathbf{h}^{(k)}_{s,a}, \omega_s)$, $s \in \mathcal{S}_{train}$ a set of N_s pairs, where $\mathbf{h}^{(k)}_{s,a}$ is a generic learned LLF vector and ω_s is the real HLF value to predict, in the training phase the regressor $r(\cdot) : \mathbb{R}^{H^{(k)}} \rightarrow \mathbb{R}$ aims at finding the hypersurface that best fits the data.

Whereas, in the test phase, generated models are used to predict the real value label $\hat{\omega}_s$ on a set of previously unseen recordings $\mathbf{v}_{q,f}$ for $q \in \mathcal{S}_{test}$. The overall descriptions for the recordings in $\mathcal{S}_{test}$ is provided as $\hat{\mathbf{y}}_{q,a} = [r_1(\mathbf{h}^{(k)}_{s,a}), ..., r_{N_D}(\mathbf{h}^{(k)}_{s,a})]$, where r_i is the model for the i-th descriptor.

Since it is not clear the correlation between LLF and HLF, in order to discover the most appropriate method, in this study we use a set of regression functions: linear regression (LR) [21]; ridge regression (RR) [21]; polynomial regression (PR) [21]; support vector regression (SVR) [16]; gradient-based boosting regression (GBR) [22]; ada boosting regression (ABR) [23].

4. EXPERIMENTS AND RESULTS

We collected recordings for 28 violins to compose the data set: thirteen historical violins (three Amati, two Guarnieri *del Gesú* and eight Stradivari) and fifteen modern violins from the collection of the Museo del Violino and of the Stradivari International School of Lutherie, based in Cremona, Italy. Instruments were played by a unique professional musician in an acoustic dry room, in order to be independent on the acoustic environment and the executions were recorded at a sample rate of 44,100 Hz and bit rate of 16 bits.

We used 20 randomly selected recordings to form the $\mathcal{S}_{train}$ and the remaining 8 to compose the test set $\mathcal{S}_{test}$. The recordings $\mathbf{x}_s$ were divided in frames $\mathbf{x}_{s,f}$ with the duration of 50 milliseconds with an overlap of 50%. We implemented a three-layers DBN ($K = 3$) with 50 neurons for each layer using *Theano* python library [24]. We used a pre-training learning rate of 10^{-6} and 100 epochs (i.e., iterations) for each layer. We averaged the learned features using the procedure introduced in section 3.1. The final corpus is 700 segments, 500 compose the training set and 200 compose the test set.

As far as HLFs are concerned, we collected annotations for each instrument by means of a questionnaire that was proposed to four professional violin makers. The testers were

Descriptor	Result	r	k	RMSE	R^2
Bright - Dark	Best	RR	1	0.13	0.15
	Worst	LR	All	0.16	-0.33
Warm Not Warm	Best	ABR	2	0.13	0.35
	Worst	LR	All	0.20	-0.58
Sweet Harsh	Best	ABR	All	0.13	0.42
	Worst	LR	All	0.22	-0.79
Full Not Full	Best	ABR	2	0.13	0.49
	Worst	LR	All	0.25	-0.76
Soft Hard	Best	ABR	3	0.09	0.56
	Worst	LR	All	0.17	-1
Deep Not Deep	Best	ABR	1	0.14	0.28
	Worst	LR	All	0.24	-1.08

Table 2. Best and worst results of the regression in the test set $\mathcal{S}_{test}$. Information about the regression function used and the layer of the DBN are also provided.

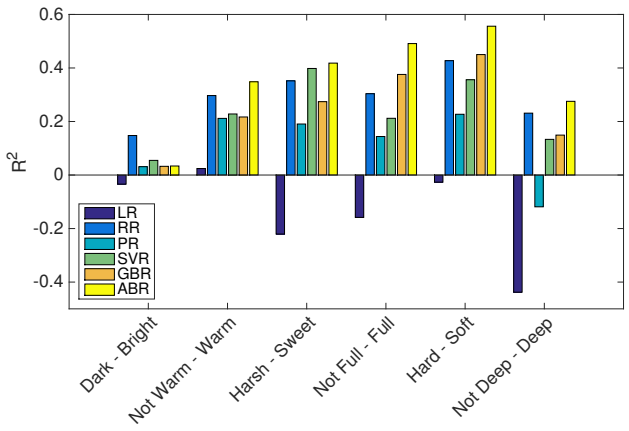

Fig. 3. Best performance for each descriptor for each prediction model in terms of R^2

asked to use the set of descriptors listed in Table 1 to annotated the instruments using a 11-point scale ranging from 0 (total prevalence of $t_i^{(l)}$) to 10 (total prevalence of $t_i^{(h)}$). The annotations were averaged over subjects and scaled between 0 and 1.

4.1. Results on automatic annotation

We evaluate the performance of the proposed regression approach in terms of $R^2 \in (-\infty, 1]$ index [21], which is a standard metric that measures the accuracy of the regression model, and in terms of Root Mean Squared Error (RMSE). Results are presented as the average of 5-fold cross-validation. Training and test sets are randomly populated. We considered either the LLF characterization provided by each layer of the DBN and the one provided by the grouping of all the layers.

For reason of space, in Table 2 we present only the most representative results which are the best and the worst performance for each descriptor. The results highlight that AdaBoost regression reached the best performance in five out of six cases, whereas the Linear regression is the worst prediction model for all the descriptors. This trend is confirmed by the Figure 3, which shows the best performance for each prediction model and descriptor. As it can be noticed, ADABoost, Gradient Boost and Ridge regression outperform the other methods. We think that is mainly due to the feature selection that these techniques embody. In fact the unsupervised approach to features learning aims at represent the input as most general as possible and a further feature selection stage, tuned on the classification task, can radically improve the performance.

From Table 2 we can also noticed that different descriptors are better modeled by a different level of the DBN: *Bright / Dark* is better modeled with the features learned in the first layer, whereas *Sweet / Harsh* descriptor is better modeled with the collection of features learned at all the levels. Intuitively, there is a connection between these results and the level of abstraction of the descriptors. *Bright / Dark*, in fact, can be easily related to the behavior of some spectral LLFs. Instead the meaning of *Sweet / Harsh* tends to be more abstract.

The different performance among the descriptors, we believe, is mainly due to the consensus of the semantics of the terms in the violin makers community. In the community, in fact, the meaning of *Deep* is not as much clear as *Hard/Soft* is. This makes the annotations less reliable and a larger dataset of recordings and annotation would be required. However, the overall accuracy of the method is promising.

5. CONCLUSIONS AND FUTURE WORKS

In this study we presented a method for modeling a set of 6 bipolar semantic descriptors for the sound quality of violins. Through a set of regression functions, the method builds a model that predicts real values (HLF) from a large set of LLFs. LLFs are learnt by exploiting the unsupervised Deep Belief Network method. The models that we obtained turn out to be quite accurate over a large set of descriptors. It is worth emphasising the fact that the learned features have been obtained through of a fully unsupervised approach and better results can be achieved with a further fine-tuning step.

Future work will be devoted to enriching the dataset of violin recordings and to exploiting other deep learning architectures, as well as fine-tuning the algorithms.

6. ACKNOWLEDGEMENTS

We are grateful to the Violin Museum Foundation and the Stradivari International School of Lutherie (both in Cremona,

Italy), for supporting the activities of timbral acquisitions on historic violins. We are particularly indebted with Fausto Cacciatori and Alessandro Voltini, for their patient work with us during the timbral acquisitions. We would also like to thank the violin makers who helped us produce the annotations and the virtuoso violinist Anastasiya Petryshak who helped us produce the audio data for the analysis.

REFERENCES

[1] J. Woodhouse, "The acoustics of the violin: a review.," *Reports on progress in physics. Physical Society (Great Britain)*, vol. 77, no. 11, pp. 115–901, Nov. 2014.

[2] M. Zanoni, F. Setragno, and A. Sarti, "The violin ontology," in *Proc. of the 9th Conference on Interdisciplinary Musicology (CIM14)*, 2014.

[3] T. Sikora H.G. Kim, N. Moreau, *MPEG-7 Audio and Beyond. Audio Content Indexing and Retrieval*, John Wiley & Sons Ltd, 2005.

[4] A. Kaminiarz and E. Lukasik, "Mpeg-7 audio spectrum basis as a signature of violin sound," in *Proc. of the 15th European Signal Processing Conference (EUSIPCO)*, 2007.

[5] J. A. Charles, D. Fitzgerald, and E. Coyleo, "Violin timbre space features," in *Proc. of the Irish Signals and Systems Conference*, 2006, pp. 471–476.

[6] E. Lukasik, "Long term cepstral coefficients for violin identification," in *Proc. of the Audio Engineering Society Convention 128 (AES128)*, 2010.

[7] A. C. Disley, D. M. Howard, and A. D. Hunt, "Timbral description of musical instruments," in *Proc. of the 9th International Conference of Music Perception and Cognition (ICMPC)*, 2006.

[8] M. Zanoni, D. Ciminieri, A. Sarti, and S. Tubaro, "Searching for dominant high-level features for music information retrieval," in *Proc. of the 20th European Signal Processing Conference (EUSIPCO)*, 2012.

[9] J. Štěpánek, "Musical sound timbre: Verbal description and dimensions," in *Proc. of the 9th International Conference on Digital Audio Effects (DAFx-06)*, 2006.

[10] R. Hirai, K. Watanabe, K. Kobayashi, and Y. Kurihara, "Measurement and evaluation of violin tone quality," in *Proc. of the SICE Annual Conference (SICE)*, 2011.

[11] Ò. Celma, P. Herrera, and X. Serra, "Bridging the music semantic gap," in *Proc. of the 1st International conference on Semantics And digital Media Technology (SAMT)*, 2006.

[12] J. Štěpánek, "Evaluation of timbre of violin tones according to selected verbal attributes," in *Proc. of the 32nd International Acoustical Conference (ICA)*, 2002.

[13] C. Fritz, A. F. Blackwell, I. Cross, J. Woodhouse, and B. C. J. Moore, "Exploring violin sound quality: investigating english timbre descriptors and correlating resynthesized acoustical modifications with perceptual properties," *The Journal of the Acoustical Society of America*, vol. 131, no. 1, pp. 783–794, 2012.

[14] A. Zacharakis, K. Pastiadis, and J. D. Reiss, "An investigation of musical timbre: uncovering salient semantic descriptors and perceptual dimensions," in *Proc. of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, 2011.

[15] Y. Yang, Y. Lin, Y. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 448–457, 2008.

[16] S. Rho and B. Han, "Svr-based music mood classification and context-based music recommendation," in *Proc. of the 17th ACM International Conference on Multimedia*, 2009.

[17] E. M. Schmidt and Y. E. Kim, "Learning emotion-based acoustic features with deep belief networks," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011.

[18] Y. Bengio, "Learning Deep Architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[19] M. Buccoli, P. Bestagini, M. Zanoni, A. Sarti, and S. Tubaro, "Unsupervised feature learning for bootleg detection using deep learning architectures," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2014.

[20] M. Buccoli, M. Zanoni, A. Sarti, and S. Tubaro, "A music search engine based on semantic text-based query," in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2013.

[21] A. Sen and M. Srivastava, *Regression analysis theory methods and applications*, Springer, New York, 1990.

[22] R. S. Zemel and T. Pitassi, "A gradient-based boosting algorithm for regression problems," in *Proc. of the 13th Conference in Neural Information Processing Systems (NIPS)*, 2000.

[23] D. P. Solomatine and D. L. Shrestha, "Adaboost.rt: a boosting algorithm for regression problems," *Proc. of the IEEE International Joint Conference on Neural Networks (IJCNN)*, 2004.

[24] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: a CPU and GPU math expression compiler," in *Python for Scientific Computing Conference (SciPy)*, 2010.