# Active learning for bird sound classification via a kernel-based extreme learning machine

Kun Qian[a),b)]
*Machine Intelligence and Signal Processing Group, Chair of Human-Machine Communication, Technische Universität München, Arcisstr. 21, Munich 80333, Germany*

Zixing Zhang and Alice Baird
*Chair of Complex and Intelligent Systems, University of Passau, Innstr. 43, Passau 94032, Germany*

Björn Schuller[a)]
*GLAM—Group on Language, Audio and Music, Department of Computing, Imperial College London, 180 Queens' Gate, Huxley Building, London SW7 2AZ, United Kingdom*

(Received 1 June 2017; revised 5 September 2017; accepted 10 September 2017; published online 4 October 2017)

In recent years, research fields, including ecology, bioacoustics, signal processing, and machine learning, have made bird sound recognition a part of their focus. This has led to significant advancements within the field of ornithology, such as improved understanding of evolution, local biodiversity, mating rituals, and even the implications and realities associated to climate change. The volume of unlabeled bird sound data is now overwhelming, and comparatively little exploration is being made into methods for how best to handle them. In this study, two active learning (AL) methods are proposed, sparse-instance-based active learning (SI-AL), and least-confidence-score-based active learning (LCS-AL), both effectively reducing the need for expert human annotation. To both of these AL paradigms, a kernel-based extreme learning machine (KELM) is then integrated, and a comparison is made to the conventional support vector machine (SVM). Experimental results demonstrate that, when the classifier capacity is improved from an unweighted average recall of 60%–80%, KELM can outperform SVM even when a limited proportion of human annotations are used from the pool of data in both cases of SI-AL (minimum 34.5% vs minimum 59.0%) and LCS-AL (minimum 17.3% vs minimum 28.4%).

## I. INTRODUCTION

Bird sounds offer a plethora of information to aid our understanding of bird mating routines and evolutionary changes.[1] Recognition of bird species via sound can make development of automated long-term bird species monitoring more feasible, and can be an effective tool for measuring the state of nature,[2] tracking climate change,[3] and assessing biodiversity within local ecosystems.[4,5]

Throughout the past two decades, ornithologists, ecologists, and engineers in both signal processing and machine learning have been working collaboratively toward applications for automatically classifying bird sound based only on audio recordings. In the early work of McIlraith and Card,[6] two-layer perceptrons were used to classify six bird species, with correct identification ranging from 82% to 93%. Somervuo *et al.*[7] studied a parametric model-based bird sound classification, and reported that average recognition accuracy for single syllables was between 40% and 50% for 14 common North-European Passerine bird species. Chen and Maher proposed a spectral peak track method that achieved 95% recognition accuracy in noisy environments from 12 natural bird species and 16 synthesized syllables.[8] Fagerlund employed mel-cepstrum parameters and low-level signal parameters within a support vector machine (SVM) classifier to improve accuracy to up to 91% and 98% for six and eight differing species, respectively.[9] Selin *et al.* introduced an encouraging method based on wavelet decomposition that achieved up to 96% accuracy for eight bird species.[10] Lee *et al.* presented a method using two-dimensional cepstral coefficients combined with Gaussian mixture models (GMMs) and vector quantization (VQ) to correctly classify nearly 84% of syllable-based units from 28 bird species.[11] The authors in Ref. 11 furthered their study by using a novel feature set based on image shape to achieve approximately up to 95% accuracy among 28 bird species.[12] An algorithm of frequency tack extraction and tonal-based features, considering background noise, were studied in Refs. 13 and 14, respectively. Briggs *et al.* proposed a multi-instance multi-label (MIML) framework in Ref. 15 for classification of multiple simultaneous bird species. Jančovič and Köküer proposed a novel method based on penalised maximum likelihood in Ref. 16. In recent years, the LifeCLEF Bird task[17] was proposed to provide the research community with a database collected from a large scale of bird species, which included 501 species within 14 027 audio recordings in 2014, extended to 1500 species

[a)] Also at: Chair of Complex and Intelligent Systems, University of Passau, Innstr. 43, Passau 94032, Germany.

[b)] Electronic mail: andykun.qian@tum.de with 36496 audio recordings in 2017. Lasseck used a method combined with large scale feature sets and segment probabilities in Ref. 18, which outperformed other submitted methods in LifeCLEF Bird task in 2014. An unsupervised feature learning method was proposed by Stowell and Plumbley,[19] which was also proven to be efficient in the BirdCLEF 2014 contest. It is noticeable that some state-of-the-art *deep learning* methods were investigated in bird sound classification and showed promising results.[20–22] In particular, it was a five-layer convolutional neural network (CNN) fed with bird sounds' spectrograms that won the BirdCLEF 2016 contest.[21]

Another direction for reducing manual annotation was to specifically focus on the detection of syllables from bird sound recordings. Kogan and Margoliash made a comparative study on the use of dynamic time warping (DTW) and hidden Markov models (HMMs) for bird song element recognition within recordings.[23] This study showed that DTW-based techniques require expert knowledge to select suitable templates, and HMMs needed more training examples than DTW templates. Another method based on evolving neural networks for unsupervised bird sound syllable classification was studied by Ranjard and Ross,[24] in which a DTW-based distance measure was designed to give an insight into the relationship of spectrogram structures between syllables. Tachibana *et al.* reported that a linear-kernel SVM can achieve around 99% recognition accuracy in day-long recordings (26055.8 ± 17672.3 syllables).[25] Tan *et al.* proposed an algorithm based on DTW and sparse representation to classify up to 81 phrase classes of *Cassin's Vireo* (*Vireo Cassinii*).[26] In their study, using only limited training data (1–5 samples per phrase), classification accuracies of 94% and 89%, respectively, on manually and automatically segmented phrases was reached. A template-based algorithm based on DTW and prominent (high-energy) time-frequency regions of training spectrograms was studied in Ref. 27; this method can outperform DTW and HMMs in most training and test conditions. In addition, this method is robust when implemented with data sets of limited sizes and within noisy background conditions.

Despite this, there are still few studies that focus on the reduction of human expert annotation for unlabeled bird sound data (segmented syllables or continuous recordings). The unavoidable truth is that expert human annotation is time-consuming, expensive, and an undesirable task for many. Previous surveys[28] have shown that, in a typical data mining project, data collection, cleaning, and annotation alone will require ∼80% of the entire time needed for the project. Specifically within the study of bird sound, there are large amounts of unlabeled audio recordings made in the field by ornithologists and amateurs, which bring forth a huge challenge for annotators. Therefore, the study of active learning (AL) for bird sound classification is significant to this domain.

In this study, two AL methods are investigated and compared, sparse-instance-based active learning (SI-AL) and least-confidence-score-based active learning (LCS-AL), which have been shown to be efficient in a preliminary study made by the authors.[29] Then, a kernel-based extreme learning machine (KELM)[30,31] is introduced, and its capacity compared with the conventionally used SVM classifier[32] when implemented in the two AL methods mentioned previously. Furthermore, a detailed comparison between algorithm efficiency and robustness will be illustrated. This article will be organised as follows: Section II will give a brief summary of related prior work. The methodology and databases used will be described in Sec. III. Experimental results are presented in Sec. IV, and the discussion in Sec. V before concluding remarks in Sec. VI.

## II. RELATION TO PRIOR WORK

Inspired by the success of AL in speech emotion recognition,[33] such methodology was introduced for bird sound.[29] To the best of our knowledge, this was the first time AL was applied for use with bird sound classification. In preliminary work on classifying 60 species of birds by sound using AL via SVM, we found that AL can reduce up to 35.2% human annotated works compared with randomly selecting samples. Extreme learning machines (ELMs)[30] were introduced first for bird sound classification in Ref. 34. ELM can outperform other conventional classifiers when fed with large scale acoustic features extracted by the openSMILE toolkit.[35] As for AL, investigated and compared were two kinds of state-of-the-art techniques, i.e., SI-AL and LCS-AL. For SI-AL, its capacity was extended for a two-class classification in Ref. 33 to multi-class classification in this study. Unlike Ref. 33, which selected samples with a medium confidence score, in this work, the samples predicted with least confidence scores were used. Combining ELM with AL is studied in Refs. 36–38. In these studies, the authors reported that ELM-based AL can be superior or at least comparable to SVM. Motivated by the success of these related works, the capacity of ELM and AL into the bird sound classification task was explored. The main contributions of this work compared with Ref. 29 are the use of an updated database of bird sounds, which includes 86 rather than 60 species of bird.[29] Second, the *confidence score* was modified through the use of "margin sampling,"[37,39] which, when estimating the trained classifier's "confidence," is more stable and efficient than the "cross entropy" used in Ref. 29. In addition, changing a new classifier is demonstrated, e.g., KELM can considerably improve the performance of AL for bird sound classification. Finally, detailed experiments on the comparison of efficiency and robustness for both SVM- and KELM-based AL algorithms are proposed.

## III. METHODOLOGY AND DATABASE

### A. Passive learning vs AL

To cope with issues of data scarcity, "passive learning" (PL) is a conventionally used method that randomly and independently selects samples from unlabeled data (see Fig. 1), asking for human experts' annotation. This method is extremely time-consuming and costly.[33] The detailed steps of PL are shown in Algorithm 1.

**Algorithm 1:** PL

```
Repeat:
(1) Randomly select $K$ samples $\phi_k$ from the unlabeled set $\Phi$
(2) Let human expert annotate the selected subset $\phi_k$
(3) Remove $\phi_k$ from the unlabeled set $\Phi$, i.e., $\Phi \leftarrow \Phi \backslash \phi_k$
(4) Add $\omega_n$ to the labeled set $\Psi$, i.e., $\Psi \leftarrow \Psi \cup \phi_k$
End: When iteration reaches a defined number, or the trained classifier
achieves a certain performance on the validation set
```

AL uses the "most informative" samples (see Fig. 1) as manual labels, and has a variety of methods to define these samples.[40] In this work, two approaches are compared, i.e., SI-AL and LCS-AL. SI-AL (see Algorithm 2) is used as a method to deal with the unbalanced distribution of this bird sound corpus. $K$ samples from data are randomly selected. Such data have been classified by a pre-trained classifier as a "sparse class" (the most informative) sample for each iteration. In LCS-AL (see Algorithm 3), the "least-confidence-score" samples ranked by a pre-trained classifier are treated as the most informative samples. It should be noted that the algorithm will select all of the candidate data (i.e., "sparse-instance" or least-confidence-score) if the number of them is less than $K$ for both Algorithms 2 and 3. Specifically, for Algorithm 2, which will randomly select, as in Algorithm 1 if there are no instances classified as sparse-instance in that iteration.

**Algorithm 2:** SI-AL

```
Repeat:
(1) Train a classifier $\mho$ based on the labeled set $\Psi$
(2) Randomly select $K$ samples $\phi_k$ from the unlabeled set $\Phi$ that are pre-
dicted by $\mho$ to be the sparse class, whose number is less than a threshold,
i.e., $K_s < K_{\max} \times sparse\_fraction$, where $K_s$ is the sample numbers of one
certain class, $K_{\max}$ is the maximum sample numbers among all data, and
$sparse\_fraction$ is a predefined ratio
(3) Let human expert annotate $\phi_k$
(4) Remove $\phi_k$ from the unlabeled set $\Phi$, i.e., $\Phi \leftarrow \Phi \backslash \phi_k$
(5) Add $\phi_k$ to the labeled set $\Psi$, i.e., $\Psi \leftarrow \Psi \cup \phi_k$
End: When iteration reaches a defined number, or the $\mho$ achieves a certain
performance on the validation set
```

**Algorithm 3:** LCS-AL

```
Repeat:
(1) Train a classifier $\mho$ based on the labeled set $\Psi$
(2) Predict the unlabeled data $\Phi$ by $\mho$, and rank the data by its prediction
confidence score
(3) Randomly select $K$ samples $\phi_k$ from the last $\vartheta$% of ranked data in $\Phi$
(4) Let human expert annotate $\phi_k$
(5) Remove $\phi_k$ from the unlabeled set $\Phi$, i.e., $\Phi \leftarrow \Phi \backslash \phi_k$
(6) Add $\phi_k$ to the labeled set $\Psi$, i.e., $\Psi \leftarrow \Psi \cup \phi_k$
End: When iteration reaches a defined number, or the $\mho$ achieves a certain
performance on the validation set
```

## B. SVM vs ELM

SVMs[32] have played an important role in classification and regression tasks for the past two decades. In theory, SVMs are given a set of training samples $\{(\mathbf{x}_i, t_i)\}_{i=1}^{n}$, where $\mathbf{x}_i \in \mathbf{R}^d$ is a feature vector in $d$-dimensional space, and $t_i$ is the predicted label of the corresponding sample. SVM aims to find the maximum margin separating hyperplane, solving the optimisation problem

$$\text{minimize}: \ L_{\text{SVM}} = \frac{1}{2}\sum_{i,j=1}^{n} t_i t_j \alpha_i \alpha_j \Omega(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^{n} \alpha_i,$$

$$\text{subject to}: \ \sum_{i=1}^{n} \alpha_i t_i = 0, \quad 0 \le \alpha_i \le C_s, \ \forall i. \tag{1}$$

Here, the $\alpha_i$ corresponds to the Lagrange multiplier of a training sample ($\mathbf{x}_i$, $t_i$), $C_s$ is a pre-defined parameter, $\Omega(\mathbf{x}_i, \mathbf{x}_j)$ is the kernel function, which could be defined as linear, polynomial, radial basis, or sigmoidal. For classification of a given test sample, a decision function is defined as

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i t_i \Omega(\mathbf{x}_i, \mathbf{x}_j) + b. \tag{2}$$

In this case, $b$ is the bias. The posterior probability of SVM estimation for a test sample can be approximated by a sigmoid function

$$P(t|f(\mathbf{x}))_{\text{SVM}} = \frac{1}{1 + \exp(A(f(\mathbf{x}) + B))}, \tag{3}$$

$A$, and $B$ is the parameter to be determined by solving a regularised maximum likelihood problem from Ref. 41.

Originally inspired by biological learning, ELMs were first proposed for single hidden layer feedforward neural networks (SLFNs).[30,31] The output function of ELM is defined as

$$f_l(\mathbf{x}) = \mathbf{h}(\mathbf{x})\boldsymbol{\beta}, \tag{4}$$

where $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_m]^T$ is the vector of output weights between the hidden layer of $m$ nodes to the $l \geq 1$ nodes of output, and $\mathbf{h}(\mathbf{x}) = [h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_m(\mathbf{x})]$ is the output row vector of the hidden layer corresponding to the input $\mathbf{x}$. ELM aims to achieve both the smallest training error and the smallest norm of output weights

$$\text{minimize}: \ \|\boldsymbol{\beta}\|_{\zeta}^{\delta_1} + C_e \|\mathbf{H}\boldsymbol{\beta} - \mathbf{T}\|_{\eta}^{\delta_2}, \tag{5}$$

where $\delta_1 > 0$, $\delta_2 > 0$, $\zeta, \eta = 0, 1/2, 2, \ldots, +\infty$, $\boldsymbol{H}$ is the hidden layer output matrix

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}(\mathbf{x}_1) \\ \vdots \\ \mathbf{h}(\mathbf{x}_n) \end{bmatrix} = \begin{bmatrix} h_1(\mathbf{x}_1) & \cdots & h_m(\mathbf{x}_1) \\ \vdots & \vdots & \vdots \\ h_1(\mathbf{x}_n) & \cdots & h_m(\mathbf{x}_n) \end{bmatrix}, \tag{6}$$

and $\mathbf{T}$ is the target (label) matrix: $\mathbf{T} = [\mathbf{t}_1^T \cdots \mathbf{t}_n^T]^T$.

In this study, KELM[31,42] is adopted, and uses a kernel matrix for ELM as follows:

$$\Omega_{\text{ELM}} = \mathbf{H}\mathbf{H}^T, \tag{7}$$

where $\Omega_{\text{ELM}(i,j)} = h(\mathbf{x}_i)h(\mathbf{x}_j)$ is a kernel function that could be defined as linear, polynomial, radial basis, or wavelet. In a multiclass case, the predicted label of a given test sample is the index number of the output node that has the highest output

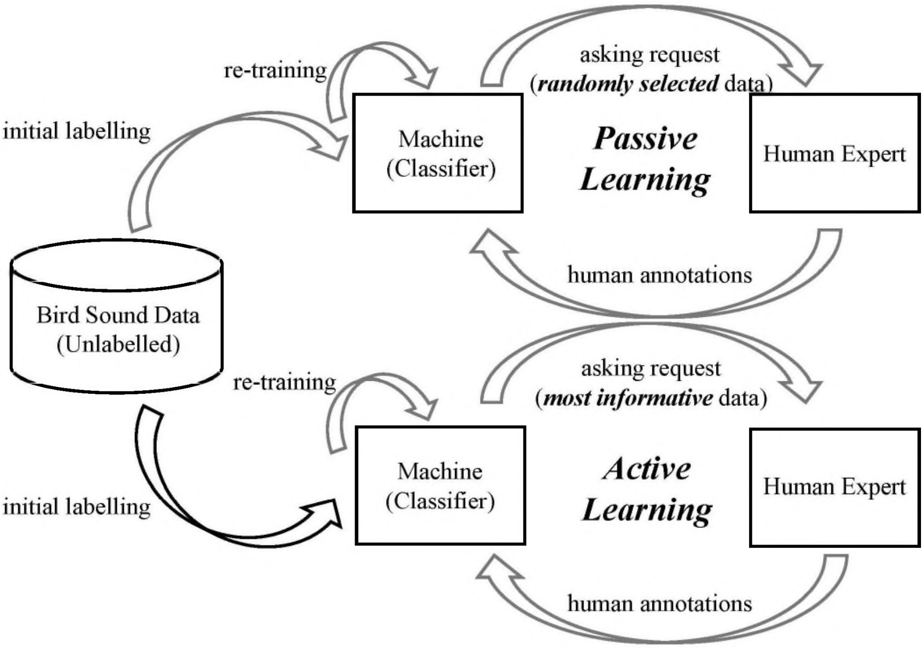

FIG. 1. The diagram of PL and AL for unlabeled bird sound data annotations. Compared with PL, AL can find the most informative unlabeled data for asking human annotations.

value corresponding to the given test sample, i.e., the predicted label of a given test sample **x** is as described in Ref. 42,

$$\text{label}(\mathbf{x}) = \underset{\xi\in\{1,2,\ldots,m\}}{\operatorname{argmax}} f_\xi(\mathbf{x}), \tag{8}$$

and the posterior probability of KELM's $P(t|f_l(\mathbf{x}))_{\text{ELM}}$ estimation can be estimated by feeding the output $f_l(\mathbf{x})$ into a "softmax" function[43] as

$$P(t|f_\xi(\mathbf{x}))_{\text{KELM}} = \frac{\exp(f_\xi(\mathbf{x}))}{\sum_{\xi=1}^{m}\exp(f_\xi(\mathbf{x}))}. \tag{9}$$

The margin sampling,[37,39] calculates the difference between the first and second largest posterior probabilities to represent the *confidence score*. The larger difference means the higher confidence score of the corresponding instance.

### C. Acoustic feature set

Motivated by the bird sound classification success in Ref. 34, the openSMILE feature extraction tool kit is used to extract large scale acoustic features for further machine learning steps. In this work, we chose the "ComParE" feature set,[44] shown to be efficient in a preliminary study.[29] The low-level descriptors (LLDs) used in ComParE feature set are listed in Table I. With functionals (refer to Ref. 45) applied to the LLDs, in total there 6373 features are extracted from bird sound data. Before feeding into the classifier, all the features extracted from the bird sound data were standardized to eliminate the effect of outliers.

TABLE I. LLDs in the ComParE feature set. RMS, root-mean-square; ZCR, zero-crossing rate; RASTA, representations relative spectra; MFCCs, mel-frequency cepstral coefficients; SHS, subharmonic summation; HNR, harmonics to noise ratio.

| |
|---|
| Group A (59) |
| Loudness, modulation loudness, RMS energy, ZCR, |
| RASTA auditory bands 1–26, MFCCs 1–14, |
| Energy 250–650 Hz, energy 1–4 kHz, |
| Spectral Roll-off Point 0.25,0.50,0.75,0.90, spectral flux, |
| entropy, variance, slope, |
| skewness and kurtosis, harmonicity, |
| sharpness (auditory), centroid (linear) |
| Group B (6) |
| $F0$ via SHS, probability of voicing, jitter (local and delta), |
| shimmer, log HNR (time domain) |

### D. Bird sound database

The bird sound data used were provided by the Museum für Naturkunde Berlin (MNB),[50] Berlin, Germany. The original database contains 273 species (subspecies) of bird sound within 6487 audio recordings. 86 species (5060 audio recordings) were chosen, which have a minimum of 20 audio recordings. The minimum, maximum, and average time duration of these recordings is 0.330 s, 59.033 s, and 2.835 s, respectively. The entire time duration of the used database is approximately 4.0 h. As shown in Table II, among each species of bird sound, 20% (≈1043) of instances were randomly selected for the validation set. To make a comparison of algorithms' efficiency and robustness, two scales of initial supervised training sets were set up, i.e., 10% and 20% from each species of bird sound, which generates 539 and 1030 instances, respectively. Finally, the rest of the data will be the unlabeled pool data set, imitating the human annotation process by feeding the real labels of the pool data to the classifiers.

TABLE II. The number of instances and percentage of total bird sound database in experiments.

| Initial | Pool | Validation | Σ |
|---|---|---|---|
| 539 (10%) | 3478 (70%) | 1043 (20%) | 5060 (100%) |
| 1030 (20%) | 2987 (60%) | 1043 (20%) | 5060 (100%) |

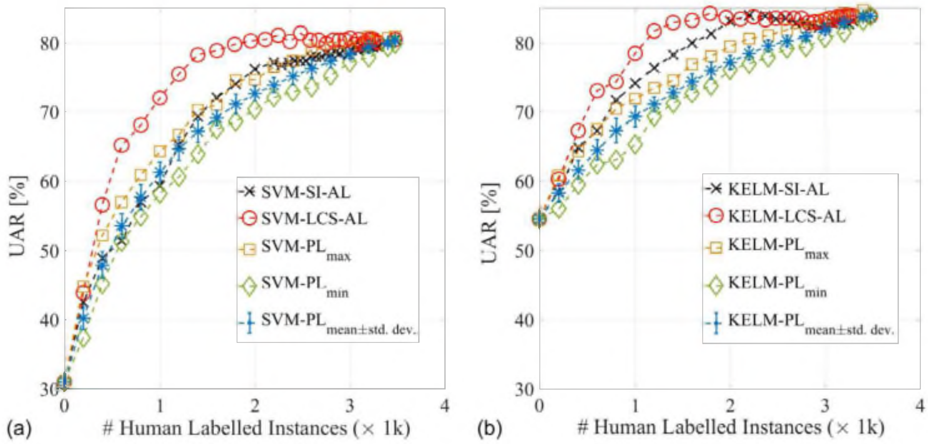

FIG. 2. (Color online) Comparison of UARs vs number of *human* labeled instances between algorithms with 539 initial supervised training instances: (a) by SVM; (b) by KELM. The measures for PL are shown with 20 independent runs.

# IV. EXPERIMENTAL RESULTS

## A. Experimental setup

The experiments on PL and AL are all implemented in the environment of MATLAB R2016b by MathWorks (Natick, MA). SVM is implemented by the popular toolkit LIBSVM[46] in a C executable environment. ELM is implemented in MATLAB. Based on a fair comparison, the kernels of both SVM and KELM are set as the *polynomial kernels* (refer to Ref. 47)

$$\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = (\gamma_* \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j + c)^d. \tag{10}$$

Here, $c$ and $d$ are set empirically to 1 and 10, respectively, by the previous preliminary experiments. The $\gamma_*$-value is an empirically defined parameter, which is set to be 1/6373 and 1 for SVM and KELM, respectively. The parameters $C_s$ and $C_e$ are set to be the same as ten by previous experiments selected from the grids of $10^{-5}, 10^{-4}, \ldots, 10^4, 10^5$. The *sparse_fraction* (mentioned in Algorithm 2) and $\vartheta\%$ (mentioned in Algorithm 3) are set to be 0.5% and 10%, respectively. All the experiments were made on the same desktop personal computer (PC) with the CPU's configuration of Intel Core™ i7–4790@3.60 GHz (Santa Clara, CA).

The evaluation metric of classification performance is the unweighted average recall (UAR),[48] which is defined as

$$\mathrm{UAR} = \frac{\sum_{\kappa=1}^{m} \mathrm{Recall}_\kappa}{m}. \tag{11}$$

Here, $m$ is the number of classes, and $\mathrm{Recall}_\kappa$ is the correct *accuracy* of the $\kappa$th class. In this study, UAR is more reasonable as the evaluation metric than *accuracy* due to the unbalanced distribution of bird species. To make an effective

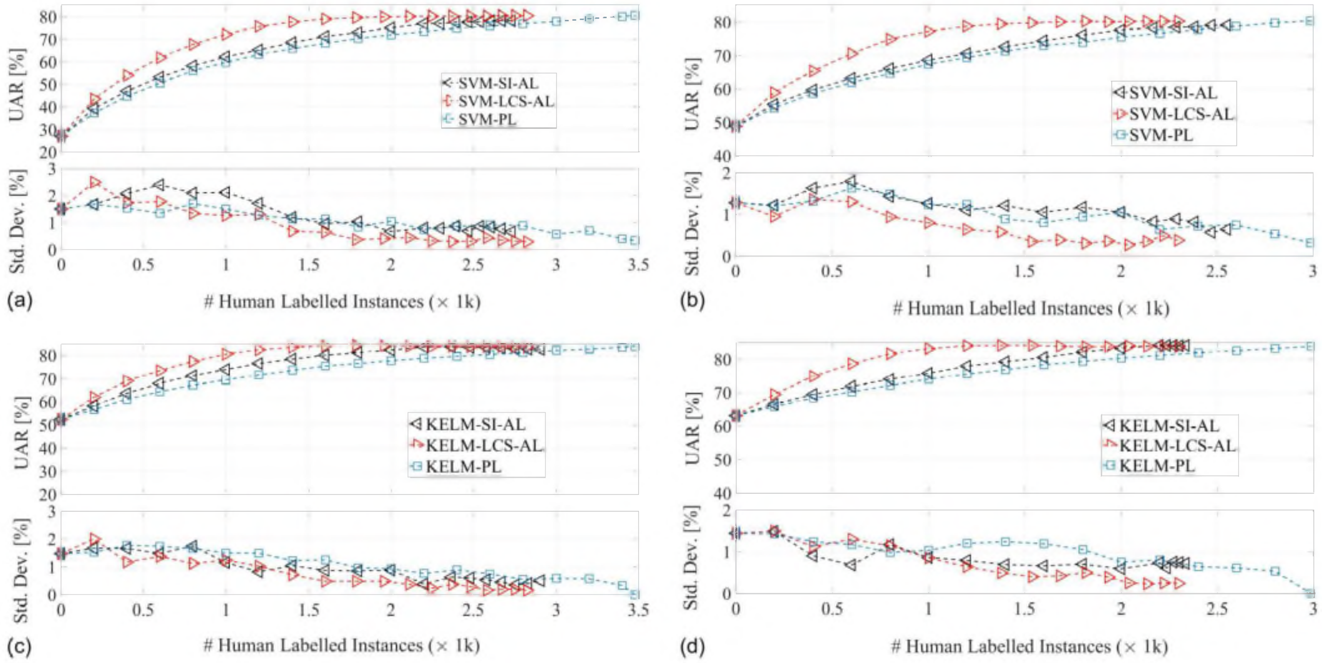

FIG. 3. (Color online) Comparison of UARs vs number of *human* labeled instances between algorithms across 20 independent runs (both for the averaged UAR and standard deviation): (a) by SVM with 539 initial supervised training instances; (b) by SVM with 1030 initial supervised training instances; (c) by KELM with 539 initial supervised training instances; (d) by KELM with 1030 initial supervised training instances. The charts only illustrate the UARs in common iterations of PL, SI-AL, and LCS-AL.

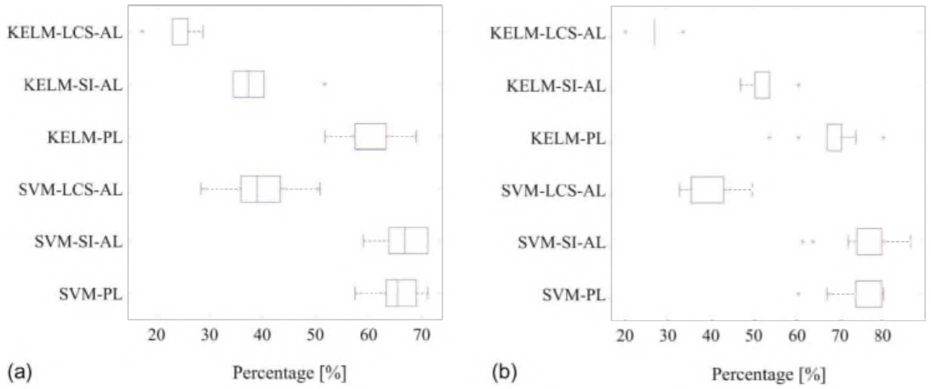

FIG. 4. (Color online) The percentage of used *human* labeled instances in total pool data when the performance (UAR) was improved from 60.0% to 80.0%: (a) with 539 initial supervised training instances; (b) with 1030 initial supervised training instances. The values are averaged across 20 independent runs.

comprehensive comparison, this study will use all of the pool data set during the final iteration of each algorithm.

### B. Comparison of PL and AL

In the experiments comparing the performance of PL and AL, the initial size was set to 539 (refer to Table II). Figure 2 shows the UARs achieved by the trained classifier vs the corresponding human labeled instances in the pool data set. For both SVM and KELM, AL is more efficient in improving the trained classifier's performance than PL (except SVM-SI-AL). In particular, LCS-AL can achieve the best recognition performance during its earlier iterations. Twenty randomly replicating experiments were run on PL; however, the maximum UARs of PL still yield to ALs (except SVM-SI-AL). It should be noted that, in these experiments, LCS-AL is superior to SI-AL for finding the most informative samples. In particular, for SVM, SI-AL cannot improve the classifier's performance compared with PL [see Fig. 2(a)]. Compared with PL, AL tends to have a slight decrease after reaching its highest point. This is due to the most informative samples being fed into the classifier; other samples bring uncertain information to the model.

### C. Comparison of robustness

To compare the robustness of different AL methods, i.e., SVM-SI-AL, SVM-LCS-AL, KELM-SI-AL, and KELM-LCS-AL, two scales were selected from the initial training set, 539 and 1030 (refer to Table II). Both scales are randomly generated and equally fed into different algorithms by 20 independent replica runs. The averaged UAR and standard deviation of the 20 independent runs are shown in Fig. 3 (results are given by common iterations among different algorithms). In such a study, LCS-AL is the fastest algorithm and can improve classifier performance at early iterations for both SVM and KELM. Compared with PL, LCS-AL can improve the classifier's performance using much less human labeled instances and with less instability (smaller standard deviation than PL). In Fig. 3, it can be seen that SI-AL is not superior to PL for SVM. However, SI-AL can reduce the number of human labeled instances, and improves the robustness when applied to KELM.

KELM performance is superior to SVM when trained by an initial data set (see Figs. 2 and 3). To make a fair comparison of SVM and KELM for their capacity to reduce human annotation work, the algorithms' performance is evaluated by observing a common range of improvement for UAR (from 60.0% to 80.0%). Figure 4 shows the percentage (in statistical box plots) of used human labeled instances in the pool data set within the UAR range from 60.0% to 80.0% by varied algorithms. KELM can outperform SVM on reducing human annotations by each corresponding learning strategy, i.e., PL, SI-AL, and LCS-AL.

## V. DISCUSSION

Tables III and IV illustrate the percentage of human labeled instances used in the pool data set when UAR was improved from 60.0% to 80.0% by each algorithm within 539 and 1030 initial training instances, respectively. We list the minimum, maximum, mean, and median values of percentage (in %) from 20 independent runs. For SVM, SI-AL was not as

TABLE III. The percentage (%) of used *human* labeled instances when the performance (UAR) was increased from 60.0% to 80.0% with 539 initial supervised training instances. The bold entries represent the best performance of each AL algorithm.

| | | Minimum | Maximum | Mean | Median |
|---|---|---|---|---|---|
| SVM | PL | 57.5 | 71.2 | 65.8 | 65.5 |
| | SI-AL | 59.0 | 71.2 | 66.8 | 66.9 |
| | LCS-AL | **28.4** | **50.9** | **40.0** | **39.0** |
| KELM | PL | 51.8 | 69.0 | 61.5 | 63.3 |
| | SI-AL | 34.5 | 51.8 | 38.0 | 37.4 |
| | LCS-AL | **17.3** | **28.8** | **23.9** | **23.0** |

TABLE IV. In this case, the initial supervised training instances were increased to 1030. The table shows percentage (%) of used *human* labeled instances when the performance (UAR) was increased from 60.0% to 80.0%. The bold entries represent the best performance of each AL algorithm.

| | | Minimum | Maximum | Mean | Median |
|---|---|---|---|---|---|
| SVM | PL | 60.3 | 80.3 | 76.7 | 79.9 |
| | SI-AL | 61.3 | 86.6 | 77.6 | 79.9 |
| | LCS-AL | **32.7** | **49.7** | **40.6** | **43.0** |
| KELM | PL | 53.6 | 80.3 | 67.0 | 67.0 |
| | SI-AL | 46.9 | 60.3 | 53.5 | 53.6 |
| | LCS-AL | **20.1** | **33.5** | **26.1** | **26.8** |

TABLE V. Significance levels of the averaged UARs obtained from the statistical comparison (one sided Student's *t*-test) between iteration: 4–18 with 539 initial supervised training instances [800–3478 instances for SVM-PL, 200–2732 (averaged) instances for SVM-SI-AL, 800–2674 (averaged) instances for SVM-LCS-AL, 800–3478 instances for KELM-PL, 200–2906 (averaged) instances for KELM-SI-AL, 800–2927 (averaged) instances for KELM-LCS-AL].

| Sign. Levels | SVM-PL | SVM-SI-AL | SVM-LCS-AL | KELM-PL | KELM-SI-AL | KELM-LCS-AL |
|---|---|---|---|---|---|---|
| SVM-PL | | | | | | |
| SVM-SI-AL | | | | | | |
| SVM-LCS-AL | ▒ | ▒ | | | | |
| KELM-PL | ▒ | ░ | | | | |
| KELM-SI-AL | █ | █ | | ░ | | |
| KELM-LCS-AL | █ | █ | █ | █ | ░ | |

░ p<.05 ▒ p<.01 █ p<.001

TABLE VI. Significance levels of the averaged UARs obtained from the statistical comparison (one sided Student's *t*-test) between iteration: 4–15 with 1030 initial supervised training instances [800–2987 instances for SVM-PL, 200–2552 (averaged) instances for SVM-SI-AL, 800–2550 (averaged) instances for SVM-LCS-AL, 800–2987 instances for KELM-PL, 200–2334 (averaged) instances for KELM-SI-AL, 800–2287 (averaged) instances for KELM-LCS-AL].

| Sign. Levels | SVM-PL | SVM-SI-AL | SVM-LCS-AL | KELM-PL | KELM-SI-AL | KELM-LCS-AL |
|---|---|---|---|---|---|---|
| SVM-PL | | | | | | |
| SVM-SI-AL | | | | | | |
| SVM-LCS-AL | ▒ | ▒ | | | | |
| KELM-PL | ▒ | ░ | | | | |
| KELM-SI-AL | █ | █ | | | | |
| KELM-LCS-AL | █ | █ | █ | █ | ░ | |

░ p<.05 ▒ p<.01 █ p<.001

superior as PL for reducing human annotations, and used even more human labeled instances than PL to improve the classifier's UAR from 60.0% to 80.0%. SVM-LCS can reduce approximately 20%–29% (within 539 initial training instances) to more than approximately 30%–35% (within 1030 initial training instances) human labeled instances from pool data set compared with PL. SI-AL works well for KELM, which reduces approximately 17%–25% to 6%–20% human annotation work by PL. KELM-LCS-AL is the best one in this study, and reduces approximately 35%–40% to approximately 33%–47% human labeled instances than PL. Finding from this that, within more initial training instances, AL can be stronger at locating the most informative samples, leading to less human labeled instances.

The significant levels (by one sided Student's *t*-test[49]) of averaged UARs by comparing different algorithms are shown in Tables V (within 539 initial training instances) and VI (within 1030 initial training instances), respectively. To eliminate the effect of early iterations' instability, one sided Student's *t*-test was set at the beginning of the fourth iteration. The comparison range will be a common length for each algorithm, i.e., the 4th to 18th iteration by each algorithm within 539 initial training instances, and the 4th to 15th iteration by each algorithm within 1030 initial training instances. The comparisons are made between a pair of differing strategies listed in the left column and the top row of each table (see Tables V and VI). The significance levels are presented by grayscale shading based on the values $p < 0.05$, $p < 0.01$, and $p < 0.001$. KELM was found to be the best algorithm among all, then comes KELM-SI-AL. SVM-LCS-AL is superior to SVM-PL and SVM-SI-AL ($p < 0.01$), yet yields more with KELM-LCS-AL.

From this study, it has been found that LCS-AL is far more efficient than SI-AL for SVM, specifically, SVM-SI-AL did not show any improvement compared with SVM-PL. This may be the reason that LCS-AL is well-matched to SVM's boundary-learning behavior.[32] In particular, a "sampling margin" has been used in LCA-AL, which focuses on distinguishing the two most similar possible classes of a given sample. Furthermore, KELM's higher performance over SVM in this study could be explained in two ways. First, when a same kernel was used, SVM tends to find a solution sub-optimal to KELM's solution.[31] Then, KELM can be directly applied into multi-class cases while SVM has to convert and indirectly solve the multi-class problems to some type of binary classification problems, which might change the application property and distribution.[31]

## VI. CONCLUSION

This study proposed a kernel-based extreme learning machine–based active learning (KELM-AL) method for bird sound classification in which we compared the KELM-based AL and SVM-based active learning (SVM-AL) to find improvement for recognition performance, and also to evaluate robustness through differing initial training sets. Experimental results showed that KELM-AL can reduce up to 47% human labeled instances, while SVM-AL can reduce up to 37% human labeled instances when compared to PL. Among AL, LCS-AL was superior to SI-AL within a one sided Student's *t*-test at $p < 0.01$ (for SVM) and $p < 0.05$ (for KELM). Future work will include the comparison of more advanced AL methods via KELM in the classification of bird sound, focusing on methods to handle such large amounts of unlabeled bird sound data.

## ACKNOWLEDGMENTS

This work was partially supported by the China Scholarship Council (CSC) and the European Union's Seventh Framework under Grant No. 338164 (ERC StG iHEARu).

## APPENDIX: BIRD SPECIES USED IN THIS STUDY

The Latin names and the instance numbers of the bird sound database used in this study are listed in Table VII.

TABLE VII. The *Latin* name and the number of instances for each bird species.

| Latin name | Instances | Latin name | Instances | Latin name | Instances |
|---|---|---|---|---|---|
| *Acrocephalus arundinaceus* | 59 | *Acrocephalus palustris* | 37 | *Acrocephalus scirpaceus* | 59 |
| *Aegolius funereus* | 50 | *Alauda arvensis* | 24 | *Anthus petrosus* | 24 |
| *Anthus pratensis* | 38 | *Anthus trivialis* | 44 | *Asio otus* | 70 |
| *Athene noctua* | 22 | *Botaurus stellaris* | 22 | *Caprimulgus europaeus* | 20 |
| *Carpodacus erythrinus* | 31 | *Certhia brachydactyla* | 42 | *Certhia familiaris* | 22 |
| *Chloris chloris* | 34 | *Chroicocephalus ridibundus* | 21 | *Corvus corax* | 21 |
| *Cyanistes caeruleus* | 36 | *Dendrocopos major* | 53 | *Dendrocopos medius* | 50 |
| *Dendrocopos minor* | 28 | *Dryocopus martius* | 53 | *Emberiza calandra* | 68 |
| *Emberiza citrinella* | 86 | *Emberiza hortulana* | 279 | *Emberiza rustica* | 32 |
| *Emberiza schoeniclus* | 119 | *Erithacus rubecula* | 26 | *Ficedula hypoleuca* | 44 |
| *Ficedula parva* | 35 | *Fringilla coelebs* | 261 | *Fringilla montifringilla* | 23 |
| *Fulica atra* | 59 | *Gallinula chloropus* | 28 | *Garrulus glandarius* | 24 |
| *Hirundo rustica* | 23 | *Jynx torquilla* | 20 | *Locustella naevia* | 23 |
| *Lophophanes cristatus* | 64 | *Lullula arborea* | 22 | *Luscinia luscinia* | 133 |
| *Luscinia megarhynchos* | 179 | *Motacilla alba* | 24 | *Muscicapa striata* | 23 |
| *Oriolus oriolus* | 23 | *Parus major* | 145 | *Passer domesticus* | 23 |
| *Passer montanus* | 22 | *Periparus ater* | 187 | *Phalacrocorax carbo* | 29 |
| *Phoenicurus ochruros* | 43 | *Phoenicurus phoenicurus* | 155 | *Phylloscopus bonelli* | 64 |
| *Phylloscopus canariensis* | 38 | *Phylloscopus collybita* | 132 | *Phylloscopus ibericus* | 129 |
| *Phylloscopus sibilatrix* | 34 | *Phylloscopus trochilus* | 133 | *Pica pica* | 20 |
| *Picus canus* | 54 | *Picus viridis* | 28 | *Podiceps cristatus* | 27 |
| *Podiceps grisegena* | 31 | *Poecile montanus* | 20 | *Porzana parva* | 23 |
| *Porzana porzana* | 66 | *Prunella modularis* | 26 | *Rallus aquaticus* | 113 |
| *Regulus ignicapilla* | 20 | *Regulus regulus* | 20 | *Saxicola rubetra* | 79 |
| *Sitta europaea* | 37 | *Strix aluco* | 34 | *Sylvia atricapilla* | 110 |
| *Sylvia borin* | 54 | *Sylvia communis* | 76 | *Sylvia curruca* | 43 |
| *Sylvia melanocephala* | 50 | *Sylvia nisoria* | 39 | *Troglodytes troglodytes* | 58 |
| *Turdus merula* | 156 | *Turdus philomelos* | 124 | *Turdus pilaris* | 54 |
| *Turdus viscivorus* | 61 | *Tyto alba* | 25 | | |

[1] C. K. Catchpole and P. J. Slater, *Bird Song: Biological Themes and Variations* (Cambridge University Press, Cambridge, UK, 2003), pp. 1–256.

[2] A. Balmford, R. E. Green, and M. Jenkins, "Measuring the changing state of nature," Trends Ecol. Evol. **18**(7), 326–330 (2003).

[3] C. Parmesan and G. Yohe, "A globally coherent fingerprint of climate change impacts across natural systems," Nature **421**(6918), 37–42 (2003).

[4] Ç. H. Şekercioğlu, G. C. Daily, and P. R. Ehrlich, "Ecosystem consequences of bird declines," Proc. Natl. Acad. Sci. U.S.A. **101**(52), 18042–18047 (2004).

[5] A. Gasc, J. Sueur, F. Jiguet, V. Devictor, P. Grandcolas, C. Burrow, M. Depraetere, and S. Pavoine, "Assessing biodiversity with sound: Do acoustic diversity indices reflect phylogenetic and functional diversities of bird communities?," Ecol. Indic. **25**, 279–287 (2013).

[6] A. L. McIlraith and H. C. Card, "Birdsong recognition using backpropagation and multivariate statistics," IEEE Trans. Signal Process. **45**(11), 2740–2748 (1997).

[7] P. Somervuo, A. Harma, and S. Fagerlund, "Parametric representations of bird sounds for automatic species recognition," IEEE Trans. Audio, Speech Lang. Process. **14**(6), 2252–2263 (2006).

[8] Z. Chen and R. C. Maher, "Semi-automatic classification of bird vocalizations using spectral peak tracks," J. Acoust. Soc. Am. **120**(5), 2974–2984 (2006).

[9] S. Fagerlund, "Bird species recognition using support vector machines," EURASIP J. Adv. Signal Process. **2007**(1), 038637 (2007).

[10] A. Selin, J. Turunen, and J. T. Tanttu, "Wavelets in recognition of bird sounds," EURASIP J. Adv. Signal Process. **2007**(1), 051806 (2007).

[11] C.-H. Lee, C.-C. Han, and C.-C. Chuang, "Automatic classification of bird species from their sounds using two-dimensional cepstral coefficients," IEEE Trans. Audio, Speech Lang. Process. **16**(8), 1541–1550 (2008).

[12] C.-H. Lee, S.-B. Hsu, J.-L. Shih, and C.-H. Chou, "Continuous birdsong recognition using Gaussian mixture modeling of image shape features," IEEE Trans. Multimedia **15**(2), 454–464 (2013).

[13] J. R. Heller and J. D. Pinezich, "Automatic recognition of harmonic bird sounds using a frequency track extraction algorithm," J. Acoust. Soc. Am. **124**(3), 1830–1837 (2008).

[14] P. Jančovič and M. Köküer, "Automatic detection and recognition of tonal bird sounds in noisy environments," EURASIP J. Adv. Signal Process. **2011**(1), 982936 (2011).

[15] F. Briggs, B. Lakshminarayanan, L. Neal, X. Z. Fern, R. Raich, S. J. Hadley, A. S. Hadley, and M. G. Betts, "Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach," J. Acoust. Soc. Am. **131**(6), 4640–4650 (2012).

[16] P. Jančovič and M. Köküer, "Acoustic recognition of multiple bird species based on penalized maximum likelihood," IEEE Signal Process. Lett. **22**(10), 1585–1589 (2015).

[17] H. Goëau, H. Glotin, W.-P. Vellinga, R. Planqué, A. Rauber, and A. Joly, "Lifeclef bird identification task 2014," in *CLEF Working Notes* (Springer, Cham, Switzerland, 2014), pp. 585–597.

[18] M. Lasseck, "Large-scale identification of birds in audio recordings," in *CLEF Working Notes* (Springer, Cham, Switzerland, 2014), pp. 643–653.

[19] D. Stowell and M. D. Plumbley, "Audio-only bird classification using unsupervised feature learning," in *CLEF Working Notes* (Springer, Cham, Switzerland, 2014), pp. 673–684.

[20] B. P. Tóth and B. Czeba, "Convolutional neural networks for large-scale bird song classification in noisy environment," in *CLEF Working Notes* (Springer, Cham, Switzerland, 2016), pp. 560–568.

[21] E. Sprengel, M. Jaggi, Y. Kilcher, and T. Hofmann, "Audio based bird species identification using deep learning techniques," in *CLEF Working Notes* (Springer, Cham, Switzerland, 2016), pp. 547–559.

[22] K. J. Piczak, "Recognizing bird species in audio recordings using deep convolutional neural networks," in *CLEF Working Notes* (Springer, Cham, Switzerland, 2016), pp. 534–543.

[23] J. A. Kogan and D. Margoliash, "Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study," J. Acoust. Soc. Am. **103**(4), 2185–2196 (1998).

[24] L. Ranjard and H. A. Ross, "Unsupervised bird song syllable classification using evolving neural networks," J. Acoust. Soc. Am. **123**(6), 4358–4368 (2008).

[25] R. O. Tachibana, N. Oosugi, and K. Okanoya, "Semi-automatic classification of birdsong elements using a linear support vector machine," PloS One **9**(3), e92584 (2014).

[26] L. N. Tan, A. Alwan, G. Kossan, M. L. Cody, and C. E. Taylor, "Dynamic time warping and sparse representation classification for birdsong phrase classification using limited training data," J. Acoust. Soc. Am. **137**(3), 1069–1080 (2015).

[27] K. Kaewtip, A. Alwan, C. O'Reilly, and C. E. Taylor, "A robust automatic birdsong phrase classification: A template-based approach," J. Acoust. Soc. Am. **140**(5), 3691–3701 (2016).

[28] D. Braha, *Data Mining for Design* and *Manufacturing*: *Methods and Applications* (Springer Science and Business Media B. V., Dordrecht, Netherlands, 2001), pp. 1–524.

[29] K. Qian, Z. Zhang, A. Baird, and B. Schuller, "Active learning for bird sounds classification," Acta Acust. Acust. **103**(3), 361–364 (2017).

[30] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," Neurocomputing **70**(1), 489–501 (2006).

[31] G.-B. Huang, "An insight into extreme learning machines: Random neurons, random features and kernels," Cognit. Comput. **6**(3), 376–390 (2014).

[32] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn. **20**(3), 273–297 (1995).

[33] Z. Zhang and B. Schuller, "Active learning by sparse instance tracking and classifier confidence in acoustic emotion recognition," in *Proc. INTERSPEECH*, ISCA, Portland, OR (2012), pp. 362–365.

[34] K. Qian, Z. Zhang, F. Ringeval, and B. Schuller, "Bird sounds classification by large scale acoustic features and extreme learning machine," in *Proc. GlobalSIP*, IEEE, Orlando, FL (2015), pp. 1317–1321.

[35] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: The Munich versatile and fast open-source audio feature extractor," in *Proc. ACM MM*, ACM, Firenze, Italy (2010), pp. 1459–1462.

[36] E. G. Horta and A. de Pádua Braga, "An extreme learning approach to active learning," in *Proc. ESANN*, Bruges, Belgium (2014), pp. 613–618.

[37] H. Yu, C. Sun, W. Yang, X. Yang, and X. Zuo, "Al-elm: One uncertainty-based active learning algorithm using extreme learning machine," Neurocomputing **166**, 140–150 (2015).

[38] Y. Zhang and M. J. Er, "Sequential active learning using meta-cognitive extreme learning machine," Neurocomputing **173**, 835–844 (2016).

[39] T. Scheffer, C. Decomain, and S. Wrobel, "Active hidden Markov models for information extraction," in *International Symposium on Intelligent Data Analysis* (Springer, Cascais, Portugal, 2001), pp. 309–318.

[40] B. Settles, "Active learning literature survey," Computer Sciences Technical Report, University of Wisconsin-Madison (2010).

[41] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," Adv. Large Margin Classifiers **10**(3), 61–74 (1999).

[42] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," IEEE Trans. on Syst., Man Cybern., Part B (Cybern.) **42**(2), 513–529 (2012).

[43] C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer, New York, 2006), pp. 115–116.

[44] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The Interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. INTERSPEECH*, Lyon, France (2013), pp. 148–152.

[45] F. Eyben, *Real-Time Speech and Music Classification by Large Audio Feature Space Extraction* (Springer International, Switzerland, 2015), pp. 229–234.

[46] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," ACM Trans. on Intell. Syst. Technol. **2**(3), 1–27 (2011), software available at https://www.csie.ntu.edu.tw~cjlin/libsvm/.

[47] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond* (MIT Press, Cambridge, MA, 2002), pp. 45–47.

[48] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 emotion challenge," in *Proc. INTERSPEECH*, Brighton, UK (2009), pp. 312–315.

[49] M. R. Spiegel, J. J. Schiller, R. A. Srinivasan, and M. LeVan, *Probability and Statistics* (McGraw-Hill, New York, 2009), pp. 213–264.

[50] http://www.animalsoundarchive.org/RefSys/Statistics.php.