Trustability-based dynamic active learning for crowdsourced labelling of emotional audio data

Simone Hantke, Alexander Abstreiter, Nicholas Cummins, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Hantke, Simone, Alexander Abstreiter, Nicholas Cummins, and Björn Schuller. 2018. "Trustability-based dynamic active learning for crowdsourced labelling of emotional audio data." *IEEE Access* 6: 42142–55. https://doi.org/10.1109/access.2018.2858931.

Trustability-Based Dynamic Active Learning for Crowdsourced Labelling of Emotional Audio Data

SIMONE HANTKE[1,2], ALEXANDER ABSTREITER[3], NICHOLAS CUMMINS[1], (Member, IEEE), AND BJÖRN SCHULLER[1,4], (Fellow, IEEE)

[1]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany
[2]Machine Intelligence & Signal Processing Group, Technische Universität München, 80333 München, Germany
[3]Chair of Complex & Intelligent Systems, University of Passau, 94032 Passau, Germany
[4]Group on Language, Audio & Music, Department of Computing, Imperial College London, London SW7 2AZ, U.K.

Corresponding author: Simone Hantke (simone.hantke@informatik.uni-augsburg.de)

This work was supported by the European Community's Seventh Framework Programme through the ERC Starting Grant iHEARu under Grant 338164.

ABSTRACT The process of collecting annotated data is expensive and time-consuming. Making use of crowdsourcing instead of experts in a laboratory setting is a viable alternative to reduce these costs. However, without adequate quality control the obtained labels may be less reliable. Whereas crowdsourcing reduces only the costs per annotation, another technique, active learning, aims at reducing the overall annotation costs by selecting the most important instances of the dataset and only asking for manual annotations for these selected samples. Herein, we investigate the advantages of combining crowdsourcing and different iterative active learning paradigms for audio data annotation. Further, we incorporate an annotator trustability score to further reduce the labelling effort needed and, at the same time, to achieve better classification results. In this context, we introduce a novel active learning algorithm, called Trustability-based dynamic active learning, which accumulates manual annotations in each step until a trustability-weighted agreement level of annotators is reached. Furthermore, we bring this approach into the real world and integrate it in our gamified intelligent crowdsourcing platform iHEARu-PLAY. Key experimental results on an emotion recognition task indicate that a considerable relative annotation cost reduction of up to 90.57 % can be achieved when compared with a non-intelligent annotation approach. Moreover, our proposed method reaches an unweighted average recall value of 73.71 %, while a conventional passive learning algorithm peaks at 60.03 %. Therefore, our novel approach not only efficiently reduces the manual annotation work load but also improves the classification performance.

INDEX TERMS Audio processing, crowdsourcing, dynamic active leaning, machine learning, user trustability.

I. INTRODUCTION

The success of modern intelligent processing systems and their underlying supervised machine learning techniques is largely owed to the availability of suitable training data. The amount and quality of this manually labelled data is a crucial step in building supervised classifiers. Typically, data annotation is performed by groups of selected experts in a controlled laboratory setting. Whilst yielding high quality labels, this annotation procedure is costly, time-consuming, and tedious work [1]–[5] which therefore leads to a scarcity of labelled data, especially in the field of speech processing. As a result, this slows down the growth and success of the development of a wide range of such systems [6].

Compared with the small amount of available labelled data, there is a wide range of unlabelled speech data available, ideally suited for the development of these systems. Present technologies, such as the Internet of Things, have made it easier than ever to collect vast, inexpensive, and truly big amounts of data. Furthermore, online sources and social media platforms like Youtube or Facebook put free and massive amounts of speech data online every minute. Nevertheless, this freely available data lacks reliable labels. In this regard, recent research projects have turned away from only gathering labels in a controlled laboratory setting and made use of crowdsourcing. Hereby, the annotation work is being outsourced to an unspecific group of people in the internet.

Crowdsourcing, provided a large enough set of annotators is used, has been shown to be a viable alternative to conventional labelling paradigms [6]–[9].

Whilst crowdsourcing has many positive aspects including efficiency and cost reduction, the online recruitment of anonymous annotators still requires a large amount of effort, since at least as many labels as there are unlabelled data instances need to be provided. Recently, several intelligent approaches have been proposed to leverage unlabelled data, one of the most promising being Active Learning [5], [10]–[12]. These state-of-the art optimisation techniques reduce the number of data instances which require manual labelling [13]–[15] and are therefore capable of reducing the time consuming and expensive manual labelling work in the first place [1], [16]–[18].

In this context, we recently developed the gamified intelligent crowdsourcing platform iHEARu-PLAY [19], [20]. The platform offers audio, video and image labelling for a diverse range of annotation tasks. Based on our initial work integrating active learning into the platform [20], we herein expanded our active learning algorithms and verify them using several emotion recognition experiments. The aim of these experiments is to identify techniques which combine the advantages of crowdsourcing and active learning to efficiently reduce the number of needed annotations.

A. RELATED WORK

The main concept of an *Active Learning* (AL) approach is based on the concept that the algorithm can improve the classification accuracy with as little training data as possible by actively choosing the data the algorithm is most certain about [18], [21], [22]. Previous research has shown that acquiring only labels for instances which the trained model cannot predict a label and therefore is most uncertain about reduces the amount of annotation costs, while achieving an equal performance and an overall cost reduction of the annotation process [6].

Due to the promising results of AL, considerable research has been done exploring this topic and it has become a rich literature source on machine learning paradigms which efficiently exploit unlabelled data for model training. AL has been applied in many diverse domains such as machine translation [1], [23], medical imaging [24], classification tasks [25], sentiment detection [13], [26], and text classification [22]. Its effectiveness was shown in multimedia retrieval [27], typical classification tasks such as automatic speech recognition [28], and speech emotion recognition [14], [29].

The main drawback of conventional (static) AL algorithms is that they still rely predominantly on annotators to provide the correct label for each instance [30]. An approach, called *Dynamic Active Learning* (DAL), tries to further reduce this fixed amount of annotators by using an adaptive query strategy without sacrificing performance [14]; however, a non-trivial amount of human intervention is needed.

Recently, these AL approaches have gained interest in combination with crowdsourcing [1], [16], [17], [20], [31]. This is not a surprising trend, AL reduces the number of instances for which manual annotations are gathered while crowdsourcing provides cheap manual labels. Lately, low-cost annotations coming from combined AL crowdsourcing tasks have been collected in different areas including machine translation, named-entity recognition, sentiment analysis, and humour classification [1], [17], [18]. Nevertheless, by relying on anonymous users within crowdsourcing, the quality and reliability of the gathered labels can eventually differ from laboratory gathered expert ones [32], [33].

Similar effects can also be observed when making use of supervised learning techniques or forms of *Semi-Supervised Learning* (SSL) used to train the classifiers [34]. Here, only a small set of labelled data is required to begin with and all the other data gets labelled by the machine afterwards. However, this can impose the disadvantage of possibly learning from wrongly labelled data through the machine and therefore the system potentially gets less accurate at every iteration. This effect was studied by many works in the literature dealing with classification tasks in the presence of label noise [35], [36] or learning with noisy labels [37], [38].

For this reasons, a mechanism to guarantee the quality and reliability of the labels is required, especially when dealing with labels acquired through crowdsourcing. It is well-known that noisy data is one of the biggest issues within crowdsourcing [39] where unreliable annotations by spammers and most importantly by malicious and carelessness users are a major known confounding issue [32], [33]. Therefore, a valid quality management mechanism needs to be setup for filtering out low quality answers to ensure a higher quality of the collected annotations.

B. CONTRIBUTIONS OF THIS WORK

In our earlier work [20], we showed the success of combining a *user trustability* property with two basic active learning query strategies in order to exploit the advantages of AL and more importantly to tackle the problem of unreliable annotations to avoid training the classifier on wrongly labelled data. In this contribution, we expand on this approach by introducing the novel *Trustability-based Dynamic Active Learning* algorithm (TDAL) (cf. Section III-F), which is a Dynamic Active Learning algorithm implementing an adaptive query strategy based on the calculated annotators' trustability scores. By addressing weakness in related algorithms, the TDAL approach ensures a high quality of the labels gathered on a crowdsourcing platform making use of detailed quality control systems (cf. Section IV).

While this novel TDAL algorithm can be integrated into a range of conventional crowdsourcing platforms, we exemplary integrated it for demonstration purposes within this work into the gamified crowdsourcing platform iHEARu-PLAY [19] to combine the advantages of both crowdsourcing and Active Learning (cf. Section V). Within the approach, an integrated trustability score for every annotator/user is

calculated, which represents how much the system trusts a user. Therefore, this score can be utilised to identify spammers, or malicious and careless users and more importantly to weight single annotations according to their reliability. The integration of the TDAL algorithm therefore aims at improving the automatic annotation process on the platform by not only increasing the reliability of the collected labels but also at reducing the annotation costs. Finally, this paper includes an in-depth analysis of the TDAL algorithm by performing an exemplary set of emotion recognition experiments (cf. Section VI and Section VII).

II. CONFIDENCE AND TRUSTABILITY CALCULATION

In this section, we introduce the employed confidence measurement method and introduce *Support Vector Machines* (SVMs) as the classification model used by the proposed algorithms. Then, we describe the novel trustability score calculation, which forms the basic concept of the proposed TDAL algorithm.

A. CONFIDENCE MEASUREMENTS

All DAL approaches actively select the data from which they learn or which require annotation. This is achieved by considering the prediction uncertainty of a trained classifier in terms of so called *confidence values* C. Such confidence measurements assess the correctness of a classification problem of an speech processing system's output. As in [34], we employ SVMs, which are a supervised learning model based on the concept of using decision hyperplanes to separate instances of different classes. This is achieved by using the decision function $f(x)$, while maximising the functional margin.

An output value of SVMs is the distance of a specific point from the separating hyperplane. Therefore, for each data instance, a confidence value is calculated by converting these distances to probability estimates within the range of [0,1] [34]. In this regard, we employ a frequently used parametric method of logistic regression [40]. For binary classification, the sigmoid function with the parameters A and B is defined as:

$$P_1(x) = \frac{1}{1 + exp(Af(x) + B)}. \tag{1}$$

$$P_0(x) = 1 - P_1(x). \tag{2}$$

The confidence value C for the predicted class is obtained by forming the difference of the posterior probabilities $P_0(x)$; $P_1(x)$ for the classes 0 and 1, respectively:

$$C(x) = |P_1(x) - P_0(x)|, \tag{3}$$

where $C(x)$ denotes the confidence value assigned to the predicted label of a given instance x. In our experiments, we consider two different query strategies, namely a *least certainty* query strategy and a *medium certainty* query strategy. Within these query strategies, the measured confidence values assigned to each instance are ranked and stored in a queue in descending order of high certainty C_h, medium certainty C_m and least certainty C_l. Accordingly, C_m represents the confidence value of the instance located in the centre of the ranking queue. Instances with least and medium confidence values are then sent to manual annotation. Formally, the query function is defined as:

$$x_c = \underset{x}{\operatorname{argmin}}\ C_x - C_{h/m/l}. \tag{4}$$

B. ANNOTATOR TRUSTABILITY CALCULATION

Within many well-known crowdsourcing platforms, such as Amazon Mechanical Turk[1] or CrowdFlower,[2] user reliability and annotation quality is usually assessed through a pretest comprised of different questions to determine if the annotator is performing the given task correctly. Inspired by this *Quality Management System* (QMS), we implemented a more detailed QMS preceding the learning algorithm to assess this novel quality mechanism called the *trustability score*.

For calculating this trustability score, we implemented several quality measuring features including consistency and control questions to compute the intra-annotator and inter-annotator agreement and integrated these measurements exemplary into the gamified crowdsourcing platform iHEARu-PLAY[3] [20].

The introduced trustability score T_u consists of three key components, namely (i) consistency questions T_{CON}, (ii) accuracy value T_{ACC}, and (iii) control questions T_{CTR}:

$$T_u = T_{CON} + T_{ACC} + T_{CTR}, \tag{5}$$

with $T_u = 100$ for instance $x = 0$ and a range [0; 100].

(i) *Consistency questions* are the percentage of repeated audio samples and function as an important factor for the trustability score calculation. Any sample that has received at least one answer from the current user can be repeated as a so called consistency question. These consistency questions are given as repeated annotation tasks to check if the user pays attention to the given task and annotates the same data instance consistently. If the given answer to such a consistency question matches the proceeding one, the consistency value T_{CON} increases and therefore the trustability score T_u. Having an answer that differs from the proceeding one, it is subtracted from the trustability score (cf. Equation 6).

$$T_{CON} = \begin{cases} T_{u(new)} = T_{u(old)} + f_t & \text{if } l_1(x) = l_2(x) \\ T_{u(new)} = T_{u(old)} - f_t & \text{otherwise,} \end{cases} \tag{6}$$

with a trustability factor f_t and a range of $T_{CON} = [0; 33]$.

Many of the well-known crowdsourcing platforms compute their annotators reliability according to this scheme (e.g., CrowdFlower using so called Test Questions[2]). However, such approaches fail to take one main aspect into account: if a user always chooses the same label for the same sample, this may be seen as consistent within the consistency measurement and the reliability of the annotator is set higher accordingly. But, if the answer itself is always not correct, this leads

[1] https://www.mturk.com
[2] https://www.crowdflower.com
[3] https://ihearu-play.eu

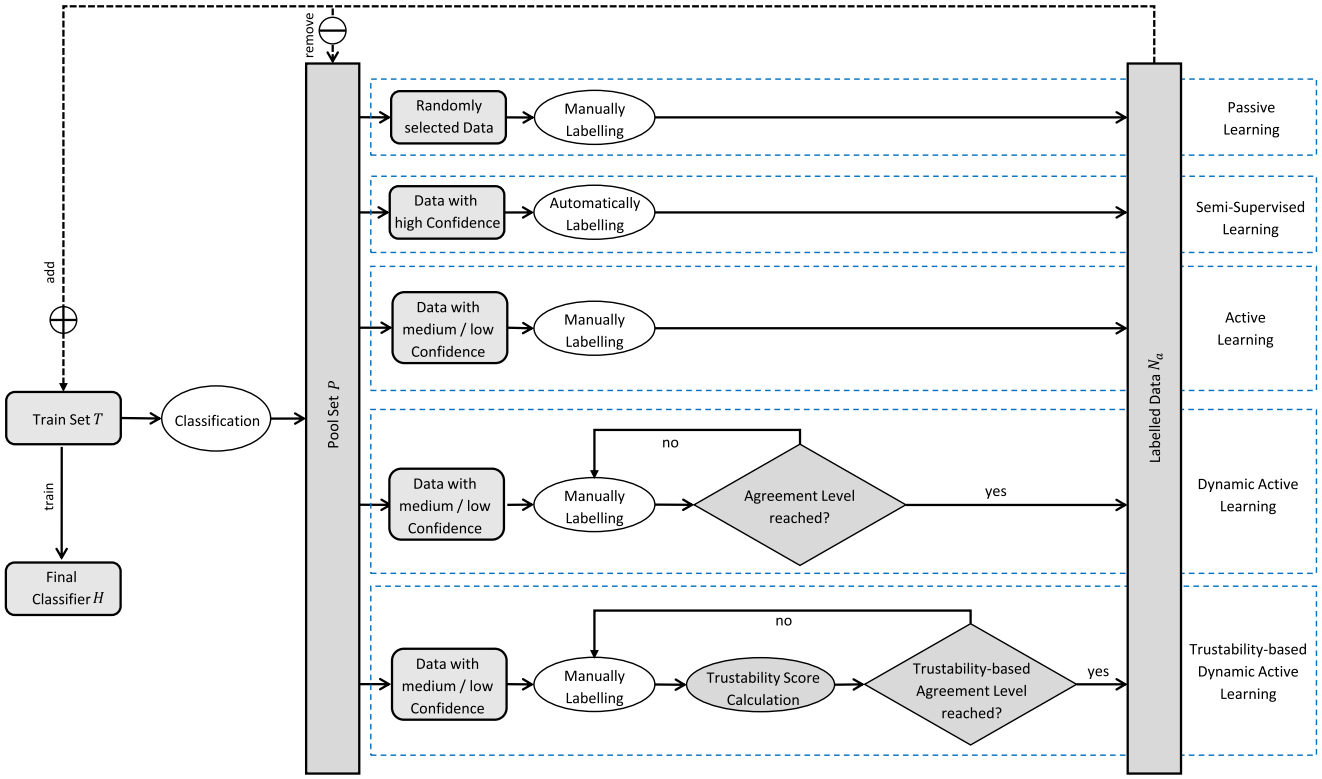

FIGURE 1. Functional diagram showing the process of the five different learning algorithms.

to false reliability results. This is a crucial gap in this measurement and is being filled in the proposed approach by our so called *accuracy value* to ensure the robustness of the trustability score measurement.

(ii) *Accuracy values* focus on the relation of the given answer to other users' answers towards the same sample. The calculation uses the difference between the current answer and the average of all other answers to check the user's deviation from the average. If the current answer is in line with the overall average answer, this annotation is considered a trustworthy one, resulting in a positive accuracy value getting added to the trustability score. If this is not the case, the annotation is most likely incorrect and the accuracy value is subtracted from the trustability score (cf. Equation 7).

$$T_{ACC} = \begin{cases} T_{u(new)} = T_{u(old)} + f_t & \text{if } l_{user}(x) = l_{all}(x) \\ T_{u(new)} = T_{u(old)} - f_t & \text{otherwise,} \end{cases} \quad (7)$$

with l_{all} being the averaged/majority label l from all users and a range of $T_{ACC} = [0; 33]$.

(iii) *Control questions*, on the other hand, contain answer possibilities which do not make sense in combination with the given task. If an annotator selects a control choice as their answer, their trustability score is decreased accordingly (cf. Equation 8).

$$T_{CTR} = \begin{cases} T_{u(new)} = T_{u(old)} + f_t & \text{if } l_{user}(x) = l_{CTR}(x) \\ T_{u(new)} = T_{u(old)} & \text{otherwise,} \end{cases} \quad (8)$$

with a range of $T_{CTR} = [0; 33]$.

Besides these three main components for calculating the trustability score, iHEARu-PLAY integrates further QMS features such as: gamification, as opposed to monetary motivation [41], pre-annotation listening checks, tracking a player's selection for highly repetitive inputs, tracking the voting time, and enforcing a minimum listening time before users can submit an answer. Our integrated QMS and the calculated trustability score represents a promising approach to obtain annotations from non-expert annotators that are qualitatively close to gold standard annotations created by experts.

III. INTELLIGENT AUDIO ANALYSIS AND ACTIVE LEARNING ALGORITHMS

In the following, we present our intelligent crowdsourcing approaches which combine different beyond state-of-the-art AL algorithms[4] with the crowdsourcing platform iHEARu-PLAY [19]. Figure 1 overviews the process of our various learning algorithms, which are going to be described in the following subsections.

A. GENERAL SETUP FOR THE LEARNING ALGORITHMS

For the learning approaches presented in this work we need a small set of already labelled data, which we can obtain either from experts or through crowdsourcing. This labelled data is split into two parts: one part is used as the training

[4]The baseline code can be found on GitHub: https://github.com/iHEARu-PLAY/iHEARu-PLAY

Algorithm 1 Passive Learning (PL)

repeat
- (Optional) Upsample training set $\mathcal{T}$ to obtain equal class distribution $\mathcal{T}_D$.
- Select subset $\mathcal{N}_a$ of pool set $\mathcal{P}$ randomly.
- Submit selected instances $\mathcal{N}_a$ to manual annotation.
- Remove $\mathcal{N}_a$ from pool set $\mathcal{P} = \mathcal{P} \setminus \mathcal{N}_a$.
- Add $\mathcal{N}_a$ together with obtained labels from the annotators to training set $\mathcal{T} = \mathcal{T} \cup \mathcal{N}_a$.

until *model training converges OR a predefined number of iterations is met.*

set for our algorithm and the other part as the test set, which evaluates the accuracy of the classifier. The unlabelled data will function as our pool set. To formalise this idea we define $\mathcal{D} = ([\boldsymbol{x}_1, l_1], \ldots, [\boldsymbol{x}_d, l_d]), i = 1, 2, \ldots d$, as a set of labelled data, where $\boldsymbol{x}_\mathrm{i}$ are the feature vectors with labels l_i. We divide this labelled data into a training set $\mathcal{T}$ and a test set $\mathcal{S}$ with $\mathcal{D} = \mathcal{T} \cup \mathcal{S}$ and $\mathcal{T} \cap \mathcal{S} = \{\}$. In addition, let $\mathcal{P} = (\boldsymbol{x}_1', \ldots, \boldsymbol{x}_u')$ be the pool set of unlabelled data with $d \ll l$.

B. PASSIVE LEARNING

As a baseline, we use a *Passive Learning* (PL) approach (cf. Algorithm 1). Within PL, the data instances which are submitted for manual labelling are simply chosen randomly. In every iteration of the algorithm, a subset $\mathcal{N}_a$ of the pool set $\mathcal{P}$ is selected randomly and submitted to the manual annotation process. After having collected the label, the data instance is removed from the pool set and added together with the gathered label to the training set $\mathcal{T}$. This procedure is repeated until all instances of the pool set are labelled.

C. SEMI-SUPERVISED LEARNING

Semi-Supervised Learning (SSL) techniques use previously labelled data to find corresponding labels for the still unlabelled data in an iterative process. In our presented experiments, a SSL variant based on the so-called *Self-Training* algorithm is used [34].

To obtain an equal class distribution in the training set $\mathcal{T}$, an upsampling step can be performed, wherein multiple copies of existing instances can be added to the classes which have a low amount of instances [42]. Afterwards, a model is trained on the existing training data $\mathcal{T}$, this is followed by the machine labelling all data in the unlabelled pool set $\mathcal{P}$ and assigning each instance a confidence score C according to Equation (3). The algorithm then chooses the set of files $\mathcal{N}_a$ with the highest confidence C_h, removes them from the pool set $\mathcal{P}$, and adds them together with the predicted label to the training set $\mathcal{T}$. Finally, the algorithm starts again from the beginning with the updated training set and repeats these steps until it reaches a certain accuracy or a maximum number of iterations (cf. Algorithm 2).

Algorithm 2 Semi-Supervised Learning (SSL) Based on the High Certainty Query Strategy [34]

repeat
- (Optional) Upsample training set $\mathcal{T}$ to obtain even class distribution $\mathcal{T}_D$.
- Use $\mathcal{T}/\mathcal{T}_D$ to train classifier H, and then classify pool set $\mathcal{P}$.
- Calculate corresponding classifier's confidence value C.
- Select subset $\mathcal{N}_a$ which contains the instances predicted with the highest confidence values C_h.
- Remove $\mathcal{N}_a$ from pool set $\mathcal{P} = \mathcal{P} \setminus \mathcal{N}_a$.
- Add $\mathcal{N}_a$ together with predicted labels to training set $\mathcal{T} = \mathcal{T} \cup \mathcal{N}_a$.

until *there is no data in the pool set predicted as belonging to the target class OR model training converges OR manual annotation is not possible.*

D. ACTIVE LEARNING

Given the highly promising results presented in our earlier work [20], we consider two basic AL algorithms with two different certainty query strategies for the classification tasks (cf. Algorithm 3). Starting with an optional upsampling procedure, an equal class distribution $\mathcal{T}_D$ in the training set $\mathcal{T}$ can be obtained. Both algorithms start by classifying all instances of the unlabelled data pool $\mathcal{P}$ using a model previously trained on the labelled data L.

Following the earlier described confidence measurement approach (cf. Section II-A), the confidence values C assigned to each instance are ranked and stored in a queue Q (in descending order). Finally, a subset N_a of $\mathcal{P}$, corresponding to those instances predicted with least and medium confidence values, are sent for manual annotation. Thenceforth, these instances are added to the training set $\mathcal{T}$ and removed from the unlabelled data set $\mathcal{P}$. This sequential process is repeated until a predefined number of instances are selected or until some stopping criterion is met [20], [34].

E. DYNAMIC ACTIVE LEARNING

The above described AL algorithms are static, meaning they wait until j manual annotations are gathered for an instance before determining the final label, using majority voting as the most popular technique. Alternately, a dynamic learning process starts by training a model on the labelled training set $\mathcal{T}$ and subsequently using this model to classify all instances of the unlabelled pool set $\mathcal{P}$. According to the least or medium certainty query strategy, a subset $\mathcal{N}_a \subset \mathcal{P}$ is selected and submitted for manual annotation. The sequential process is repeated until a certain number of instances are annotated and a predefined agreement level j is reached for every of these instances.

The main improvement of this technique compared to the static AL method is that these Dynamic Active Learning (DAL) algorithms are based on an adaptive query

Algorithm 3 Active Learning (AL) With Least and Medium Certainty Query Strategy for Classification Procedures; Adapted From [43]

repeat
  (Optional) Upsample training set $\mathcal{T}$ to obtain even class distribution $\mathcal{T}_D$.
  Use $\mathcal{T}/\mathcal{T}_D$ to train a classifier $\mathcal{H}$, then classify pool set $\mathcal{P}$.
  Calculate the corresponding classifier's confidence value C.
  Rank data based on the prediction confidence values C and store them in a queue Q.
  Choose a query strategy:
    – Least certainty query strategy: Select subset $\mathcal{N}_a$ of pool set $\mathcal{P}$ whose elements are 'at the bottom' of the ranking queue Q.
    – Medium certainty query strategy: Select subset $\mathcal{N}_a$ of pool set $\mathcal{P}$ whose elements are 'in the middle' of the ranking queue Q.
  Submit selected instances $\mathcal{N}_a$ to manual annotation.
  repeat
    Remove $\mathcal{N}_a$ from the unlabelled pool set $\mathcal{P}$, $\mathcal{P} = \mathcal{P} - \mathcal{N}_a$.
    Add $\mathcal{N}_a$ and their aggregated labels to the training set $\mathcal{T}$, $\mathcal{T} = \mathcal{T} \cup \mathcal{N}_a$.
  until
until *there is no data in the pool set predicted as belonging to the target class OR model training converges OR manual annotation is not possible OR a predefined number of iterations is met.*

strategy [34]. In this context, the DAL approach first requests a small number of annotations for every instance in the subset $\mathcal{N}_a$, and then only requests further annotations if the predefined agreement level j for one class has not been reached.

Algorithm 4 presents the pseudocode description of the DAL algorithm. If all annotators have voted for the same label l_m, the algorithm stops asking for further annotations. If the first annotators have not agreed on one label, one more annotation is requested and it is checked again if the predefined agreement level has been reached. This whole process is repeated until the predefined agreement level j is reached.

F. TRUSTABILITY-BASED DYNAMIC ACTIVE LEARNING

As previously mentioned, a major issue when gathering annotations, especially using crowdsourcing, is that the quality of the annotations can be low. This can result in training the model using wrongly labelled data, which in turn, can cause reductions in the accuracy of a classifier trained using this data. With the aim of overcoming this issue, we now introduce a novel DAL algorithm which explicitly incorporates annotator trustability-based agreement levels.

Algorithm 4 Dynamic Active Learning (DAL) With Least and Medium Certainty Query Strategy for Classification Procedures; Adapted From [34]

repeat
  (Optional) Upsample training set $\mathcal{T}$ to obtain even class distribution $\mathcal{T}_D$.
  Use $\mathcal{T}/\mathcal{T}_D$ to train a classifier $\mathcal{H}$, then classify pool set $\mathcal{P}$.
  Calculate the corresponding classifier's confidence value C.
  Rank data based on the prediction confidence values C and store them in a queue Q.
  Choose a query strategy:
    – Least certainty query strategy: Select subset $\mathcal{N}_a$ of pool set $\mathcal{P}$ whose elements are 'at the bottom' of the ranking queue Q.
    – Medium certainty query strategy: Select subset $\mathcal{N}_a$ of pool set $\mathcal{P}$ whose elements are 'in the middle' of the ranking queue Q.
  For each instance x in $\mathcal{N}_a$:
  repeat
    Submit x to the first u annotators.
    Let v be the number of votes of the label with the most votes.
    If $v >= j$: STOP
    Else repeat: select one annotator for annotation.
  until *until agreement level j is achieved*
  Remove $\mathcal{N}_a$ from pool set $\mathcal{P} = \mathcal{P} \setminus \mathcal{N}_a$.
  Add $\mathcal{N}_a$ together with obtained labels from the annotators to training set $\mathcal{T} = \mathcal{T} \cup \mathcal{N}_a$.
until *there is no data in the pool set predicted as belonging to the target class OR model training converges OR manual annotation is not possible OR a predefined number of iterations is met.*

The proposed algorithm aims at collecting only highly reliable labels, while at the same time preventing the acquisition of unnecessary annotations which do not bring further improvement to the models. In contrast to the earlier introduced DAL method (cf. Section III-E) [14], the trustability-based agreement level j is computed by using the trustability score T_u of an annotator to determine how many subsequent annotations have to be collected. In this context, the number of annotations collected for each sample depends directly on the number of trusted annotators previously queried.

The advantage of the agreement level in this approach over the AL method is that it does not stop when a pre-defined number of annotations for one data instance have been gathered. Instead, the procedure is repeated until the defined user trustability sum of the answers reaches the agreement level. By gathering annotations until it has acquired enough high-quality annotations and using the trustability-weighted majority voting (cf. Section IV), the algorithm helps to ensure a high quality of the final labels.

Algorithm 5 Trustability-Based Dynamic Active Learning With Least and Medium Certainty Query Strategy for Classification Procedures

1 **repeat**
2 (Optional) Upsample training set $\mathcal{T}$ to obtain even class distribution $\mathcal{T}_D$.
3 Use $\mathcal{T}/\mathcal{T}_D$ to train a classifier $\mathcal{H}$, then classify pool set $\mathcal{P}$.
4 Rank data based on the prediction confidence values C and store them in queue Q.
5 Choose a query strategy:
6 – Least certainty query strategy: Select subset $\mathcal{N}_a$ whose elements are 'at the bottom' of the ranking queue Q.
7 – Medium certainty query strategy: Select subset $\mathcal{N}_a$ whose elements are 'in the middle' of the ranking queue Q.
8 For each instance x in $\mathcal{N}_a$:
9 **repeat**
10 Submit x to all annotators.
11 Wait until trustability-based agreement level $\sum_{a_u \in \mathcal{A}_{xl}} (T_u + a_t) \geq j$ is fulfilled.
12 **until**
13 Remove $\mathcal{N}_a$ from pool set $\mathcal{P} = \mathcal{P} \setminus \mathcal{N}_a$.
14 Add $\mathcal{N}_a$ together with obtained labels from the annotators to training set $\mathcal{T} = \mathcal{T} \cup \mathcal{N}_a$.
15 **until** *there is no data in the pool set predicted as belonging to the target class OR model training converges OR manual annotation is not possible OR a predefined number of iterations is met.*

$\mathcal{A}_{xl}$ are all annotations in which the instance x was labelled with the label l; T_u being the user trustability; a_t being the anti-trustability weight, determining how strongly T_u should be weighted.

Algorithm 5 gives the pseudocode description of the novel proposed TDAL algorithm. Step five shows the main improvement over the DAL algorithm. First of all, the files of the chosen subset $\mathcal{N}_a$ are submitted to all annotators, instead of just to the first u users and can be dynamically removed from the pool of instances open for user annotations when the agreement level is reached. Let $\mathcal{A}_{xl} = \{a_u \in \mathcal{A}_x : a = l\}$ be the set of all annotations with label l for sample x; T_u being the trustability of user u and a_t the anti-trustability weight, determining how strongly the user trustability should be weighted. Each instance is then annotated until the following condition is fulfilled:

$$\sum_{a_u \in \mathcal{A}_{xl}} (T_u + a_t) \geq j. \tag{9}$$

Having collected enough high-quality annotations for one data instance, it is removed from the pool of instances that are available for annotation and is added to the training set together with its final label. This is repeated until all instances of $\mathcal{N}_a$ achieve the desired agreement level.

Although users with low trustability scores have a low influence on fulfilling this condition, the files are also submitted to these users in order to give them the opportunity to improve their trustability score by answering the subsequent consistency or control questions in an agreement-based manner (cf. Section II-B).

IV. TRUSTABILITY-BASED VOTING METHODS

All previously described AL algorithms acquire single-user manual annotations in every iteration for all files of the selected subset. To prevent wrongly labelled data being added to the training set, it is necessary to determine the final label of an instance out of all its single user annotations.

In this context, we assume n gathered manual annotations $\mathcal{A}_x = \{a_u : \text{user } u \text{ labelled instance } x \text{ with label } a \in \mathcal{L}\}$ for instance x with $\mathcal{L}$ being the set of all available labels. $\mathcal{A}_{xl} = \{a_u \in \mathcal{A}_x : a = l\}$ denotes all annotations in which the instance x was labelled with the label l. Given these annotations, it is the aim to determine the correct final label l_f for x.

Different annotation models have previously been introduced in the literature to merge the labels collected from the different annotators, a conventional method being majority voting, which can be seen as the pseudo-standard model and has been used for a wide range of annotation tasks [16], [17], [44]. Based on this methodology, different expanded models such as ZenCrowd [45], GLAD [46], CUBAM [47] and CrowdSynth [48] have been introduced. These methods take into account an annotators area of expertise and/or their interpretation or assumptions that each data instance has its own inherent difficulty to label. Nevertheless, given these attributes, these models should not be regarded as a neutral way of merging different labels. Within this work, we use majority voting to keep these parameters to a minimum to fully evaluate the advantages of our proposed algorithm. Therefore, we integrated the trustability score into a majority voting method for discrete sets of available labels and a median voting method for non-discrete sets.

A. MAJORITY VOTING

Majority voting is a widely used method by crowdsourcing systems to determine the final label l_f for an instance [16], [17], [44]. The main idea is to select the answer that has been chosen most often and it can be formalised as follows:

$$l_f = \underset{l \in \mathcal{L}}{\operatorname{argmax}} |\mathcal{A}_{xl}|. \tag{10}$$

B. MEDIAN VOTING

For non-discrete numeric label sets, the median of the different labels will be applied to determine the final label. The advantage of the median compared to the arithmetic mean is its robustness against extreme values [49]. In a first step all annotations are sorted according to their label value in a zero-based (start the indexing at 0) list V. Then, two different scenarios apply, one for n being odd, and the other one for n

being even, which determines the final label as follows:

- n is odd: $l_f = V[\lfloor \frac{n}{2} \rfloor]$
- n is even: $l_f = \frac{V[\frac{n}{2}]+V[\frac{n}{2}-1]}{2}$.

C. TRUSTABILITY-WEIGHTED MAJORITY VOTING

We propose a novel method making use of a conventional majority voting approach which incorporates the trustability of each user in order to achieve a more accurate final label. A score for each label will be calculated over the votes of the user in combination with their trustability T_u and an anti-trustability weight a_t. A higher weight leads to a less influence of the trustability on reaching the agreement level and computing the label. The final label l_f is therefore calculated as follows:

$$l_f = \underset{l \in \mathcal{L}}{\operatorname{argmax}} \sum_{a_u \in \mathcal{A}_{xl}} (T_u + a_t). \tag{11}$$

D. TRUSTABILITY-WEIGHTED MEDIAN VOTING

To reduce the influence of unreliable labels from users with low trustability, we propose a novel trustability-weighted median voting method. This method sorts the annotations according to the obtained value in a list V. Then, for every index m in the list, the sum of the trustabilities of the annotations $V[i]$, $0 \leq i < m$ and the sum of the trustabilities of the annotations $V[j]$, $m+1 \leq j < n-1$ are computed. After that, the index z is chosen in such a way that the difference of the two sums is as small as possible. The element at this smallest index is the final label for this instance. This proposed idea can be formalised as follows:

$$l_f = V[z],\ z = \underset{0 \leq m < n-1}{\operatorname{argmin}} \left(\sum_{i=0}^{m-1} (T_{V[i]} + a_t) - \left(\sum_{j=m+1}^{n-1} (T_{V[j]} + a_t) \right) \right). \tag{12}$$

V. INTELLIGENT CROWDSOURCING PLATFORM

For the planned evaluation, we integrate the proposed algorithms (cf. Figure 1) into our crowdsourcing platform iHEARu-PLAY which is unique, in that it provides volunteers a game-like environment in which to perform annotation tasks [19]. This gamification provides users with an intrinsic motivation which can improve the quality of the annotations and can increase the amount of user activity [50]. More importantly, the platform also has different mechanisms to ensure the reliability of the labelled data by the players via a quality control system and methods of identifying the within-user-agreement (cf. Section II-B) [20]. These scores can be used in a variety of methods in order to improve the overall annotation quality.

For the integration of the proposed (T)DAL algorithms into iHEARu-PLAY, we applied a *Support-Vector-Machine* (SVM) to train the model and its confidence C is used to determine the certainty for labelling the data files (cf. Section II-A). Furthermore, the earlier described agreement level j for a label l has to be predefined. Whenever a file reaches the agreement level, it is immediately removed from the pool set $\mathcal{P}$ and the label is computed from the individual annotations incorporating the user trustability T_u. In addition, the number of files chosen in each AL iteration has to be set before starting the process, taking into account that an extremely low value can result in the algorithm missing high certainty samples and a high value might cause the machine to label files with a medium or low certainty. If SSL is used, the instances with the highest certainty of predicting the label correctly are determined, removed from the pool set $\mathcal{P}$, and added together with the predicted label to the training set $\mathcal{T}$.

After providing all the required initial data, the system starts with executing the AL algorithm and automatically offers the chosen files for manual annotation on the platform. Calculating the trustability score of a user after every annotation, it is feasible to remove the trustability-weighted agreement level reached file from the offered data for annotators dynamically. The use of this feature minimises the collection of unneeded annotations, saving time of the user and researcher.

VI. EXPERIMENTS

The following section will overview the experimental setup and will provide all key settings required for the evaluation of the proposed machine learning algorithms.

A. TASKS

We evaluated the proposed TDAL algorithm by conducting three experiments comparing the novel TDAL with agreement level $j \in \{1, 1.5\}$, the DAL with $j \in \{2, 3\}$ and the baseline PL approach. First, we set our baseline and run the basic version of DAL and TDAL. Then, the experiment is repeated exploiting DAL and TDAL with an random upsampling step, adding multiple copies of the existing instances to the classes with a low amount of instances. Finally, we carry out a further experiment combining SSL with the upsampled DAL and TDAL approaches.

B. DATASET

For all presented experiments we used the FAU Aibo Emotion Corpus [51], which was part of the INTERSPEECH 2009 Emotion Challenge [52] and contains more than 18k instances and a total of roughly 8 hours of speech data of children playing with Sony's pet robot Aibo. The language of the recorded children is rich in emotions, because they had been instructed to control the robot with their voice, whereas in reality it was controlled in a wizard-of-oz scenario by a human operator, who was sometimes deliberately disobeyed the instructions of the children. The data was recorded at two different schools; the data of one school is used for the training and pool set and the data of the other school for the test set to ensure speaker independence and different room acoustics.

C. ACOUSTIC FEATURES

Using the in iHEARu-PLAY integrated openSMILE toolkit [53], we automatically extracted the features

TABLE 1. IS09 acoustic feature set: 16 provided low-level descriptors(LLD) and 12 Functionals.

LLD (Δ)	Functionals
ZCR	mean
RMS Energy	standard deviation energy
F0	kurtosis, skeweness
HNR	extremes: value, rel. position, range
HFCC 1-12	linear regression: offset, slope, MSE

according to the INTERSPEECH 2009 Emotion Challenge feature set [52], which was explicitly designed to have a high level of robustness for human emotion recognition and has been used successfully in a range of different kind of emotion tasks [54]–[56]. This results in a 384-dimensional feature space based on 16 frame-wise low-level-descriptors and twelve functionals applied on a per-chunk level (cf. Table 1).

D. EXPERIMENTAL SETUP

Having a mature theoretical foundation [57] and having been used to set suitable baseline results in earlier similar work [20], we used the open-source SVM implementations of LibSVM [58] for the WEKA toolkit [59]. We implemented linear kernel SVMs with a complexity of 0.1 and with *Sequential Minimal Optimization* (SMO) to improve the robustness for high dimensional feature spaces.

Initially, we trained our model with 200 randomly selected instances, resulting in a training set $\mathcal{T}$ with 1.1 % of all data. 53.3 % of the data formed the unlabelled pool set $\mathcal{P}$ and 45,6 % of the data formed the test set $\mathcal{S}$. In each iteration of the different algorithms a subset $\mathcal{N}_a$ is selected according to the applied strategy with $|\mathcal{N}_a| = 200$. This subset is then submitted to manual annotation within iHEARu-PLAY.

We stopped our learning algorithms (PL, DAL and TDAL) as soon as they have reached their maximum UAR result. In addition, we repeat this process in 20 independent runs in order to reduce the statistical outliers impacting on the results.

E. EVALUATION

1) ANNOTATIONS

For the performed experiments, we labelled the FAU Aibo Emotion Corpus with the help of the proposed algorithms taking into account the trustability of the user. We had twelve annotators (three female and nine male) between 20 and 27 years old, excluding six users who did not reveal their age. The resulting mean is 23 years with a standard deviation of 2.6 years.

The 12 annotators labelled the emotional audio files into the emotions *motherese*, *touchy*, *surprised*, *neutral*, *joyful*, *emphatic*, *angry*, *helpless*, *bored* and *other* as was originally performed in the INTERSPEECH 2009 Emotion Challenge [52]. To be able to test different agreement levels of the annotators, we adapted the proposed binary emotion classes in [52]; NEG(ative) which includes all negative emotions (angry, touchy, and emphatic), and IDL(e), containing all other emotions. Table 2 shows the frequencies of the two classes NEG and IDL for the pool and test set.

TABLE 2. Distribution of instances per class. IDL: positive and neutral emotions, NEG: negative emotions, Dur.: duration of audio files in hours.

	NEG	IDL	$\sum$	Dur. [h]
Pool	1620	8233	9853	3.95
Test	1497	6760	8257	3.90

2) BASELINE

The PL algorithm, presented in section III-B, is chosen to investigate the effectiveness of the AL algorithms. It selects the instances for manual annotations randomly and therefore can be considered as a 'non-intelligent' crowdsourcing approach.

3) MEASUREMENT

To evaluate the performance of the algorithms, the trained algorithms classify the emotion of the instances in the test set, and these labels are then compared to the labels which have been already obtained on iHEARu-PLAY. Following the recommendations in [52] and [60], we use the *Unweighted Average Recall* (UAR) to determine the classification performance. The main advantage of the UAR over other metrics such as the weighted average recall is that in an unbalanced class scenario the latter is very biased towards performance in the the biggest class, whereas the UAR tackle this problem by the inclusion of a weighting factor $1/N$, with N being the number of classes [60].

VII. RESULTS

In the following, we evaluate our proposed TDAL, comparing the performance to the baseline PL algorithm and the DAL approach (cf. Figure 1). We first perform pure AL experiments for all algorithms, followed by adding a random upsampling step into the AL procedure and concluding it with a third experiment taking a SSL step into account.

A. ACTIVE LEARNING

Firstly, in order to study the effectiveness of the TDAL method, we compare the pure DAL approach and the pure TDAL algorithm to a PL approach. Assigning our DAL an agreement level $j = 2$, the algorithm stops gathering annotations for every file after it has obtained two labels for one class and adds it together with this obtained label to the training set. The TDAL approach, on the other hand, waits until it has reached a trustability sum of $j = 1$, which resulted in this experiment in an average of 1.26 manual annotations per instance.

The results, as presented in Figure 2(a), demonstrate that both approaches outperform our achieved PL baseline UAR of 60.03 %; the DAL approach achieved a maximum UAR of 61.41 % while the TDAL algorithm gained a maximum UAR of 62.66 %.

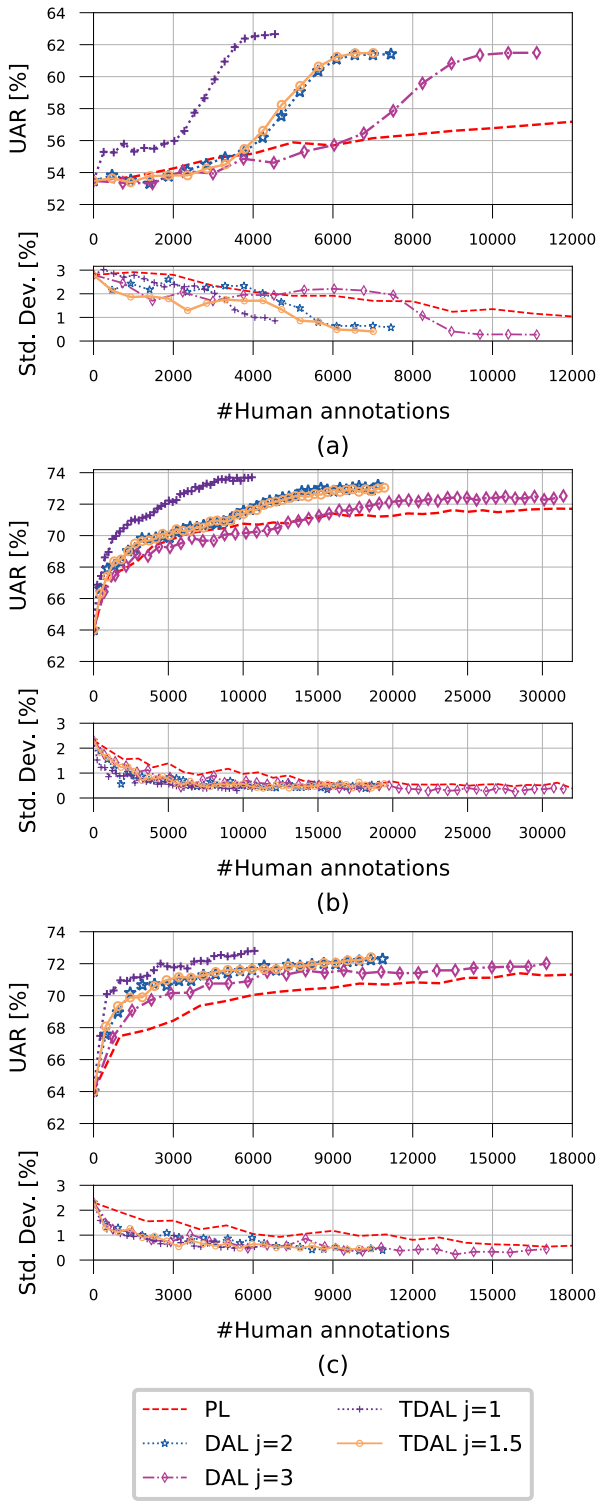

FIGURE 2. Comparison of the different (Trustability-based) Dynamic Active Learning algorithms with different agreement levels $j = 1$, $j = 1.5$, and $j = 2$ and the Passive Learning algorithm as the baseline. Average UAR and number of manual annotations are measured across 20 independent runs of each algorithm on the FAU AEC database. Results of pure AL algorithms are shown in (a), of upsampled AL in (b), and of upsampled AL with SSL in (c).

The maximum UAR, the number of annotations needed to reach this maximum and the relative annotation cost reduction of reaching the maximum UAR compared to PL are given in Table 3. As can be seen, the baseline PL algorithm collects more than 48 k annotations to achieve this maximum UAR, while the TDAL stops after only 4 549 manual labelled instances, resulting in the highest overall relative annotation cost reduction of 90.57 %. This corresponds to 3.58 hours of the total 8 hours of audio data which, when using TDAL, does not need to be annotated. Inspecting DAL with $j = 3$ and TDAL with $j = 1.5$, it can be seen that the DAL and TDAL show a similar performance (DAL 61.50% UAR, TDAL 61.48 % UAR), but the TDAL clearly outperforms the DAL stopping after 7012 annotations compared to 11 108 manual labels.

B. UPSAMPLED ACTIVE LEARNING

Next, we performed experiments making use of upsampling the under represented class before training the model. In general, PL as well as DAL and TDAL achieved better performances than without upsampling and both AL methods clearly outperform PL approach (see Figure 2(b)).

In this testing scenario, the PL approach reached a maximum UAR of 72.08 %; however, more than 48k annotations had to be collected to obtain this score. As presented in Table 3, DAL and TDAL considerably reduce the annotation load and therefore outperform the baseline. While the best DAL ($j = 2$) algorithm needs 19 014 manual annotations to reach its maximum UAR of 73.2 %, the best TDAL approach ($j = 1$) gathers only 10 584 manual annotations and reaches even a slightly higher UAR of 73.71 %. The process results in a maximum relative cost reduction compared to PL of up to 59.77 % for the DAL algorithm ($j = 2$). The TDAL method ($j = 1$) outperforms all by achieving the maximum UAR 73.71 %, while saving up to 78.07 % of the annotation effort.

C. UPSAMPLED SEMI-SUPERVISED ACTIVE LEARNING

It has been shown that further improvements in the system accuracy can be obtained by applying a SSL step after each AL step [6]. As shown in Figure 2(c), the DAL as well as the TDAL algorithms clearly outperform the PL method. All approaches have a steep learning curve in the first iterations and reach their maximum UARs of 72.29 % for DAL ($j = 2$) and 72.80 % for TDAL ($j = 1$) when every instance of the pool set has been labelled either by humans or by the machine. As shown in Table 3, all algorithms in combination with SSL further reduce the costs of the annotation process. The DAL algorithm achieves a maximum cost reduction of 77.48 %, while TDAL leads up to even 87.46 % total cost reduction, only needing 6 051 annotations.

D. DISCUSSION OF RESULTS

The previous section overviewed our achieved UAR measures and annotation cost reductions of the different proposed learning algorithms. For all approaches, the sequential addition of manual-labelled instances to an initial training set (200 per iteration) led to a continuous improvement in classification performance. Further, for all algorithms the UAR first

TABLE 3. Numeric results of twenty independent runs each on the FAU Aibo dataset. j is the agreement level of the algorithms. $\mathrm{UAR_{max}}$ denotes the maximum achieved UAR of the algorithms and their standard deviations $Std.Dev.$ [%]. $\mathrm{NA_{max.\ UAR}}$ is the number of annotations needed to achieve this $\mathrm{UAR_{max}}$. CR is the relative cost reduction when achieving the maximum compared to Passive Learning and TR is the duration of the audio files for which the annotation costs can be saved.

Algorithm	j	$\mathrm{UAR_{max}}$ [%]	$Std.Dev.$ [%]	$\mathrm{NA_{max.\ UAR}}$	CR [%]	TR [h]
Active Learning						
PL	-	60.03	0.23	48 265	-	-
DAL	2	61.41	0.57	7 453	84.56	3.34
DAL	3	61.50	0.27	11 108	76.99	3.04
TDAL	1	**62.66**	0.86	**4 549**	**90.57**	**3.58**
TDAL	1.5	61.48	0.40	7 012	85.47	3.38
Upsampled Active Learning						
PL	-	72.08	0.28	48 265	-	-
DAL	2	73.20	0.52	19 014	59.77	2.36
DAL	3	72.52	0.36	31 404	33.92	1.38
TDAL	1	**73.71**	0.38	**10 584**	**78.07**	**3.09**
TDAL	1.5	73.04	0.57	19 434	59.73	2.36
Upsampled Active Learning with Semi-Supervised Learning						
PL	-	72.08	0.28	48 265	-	-
DAL	2	72.29	0.39	10 867	77.48	3.06
DAL	3	72.02	0.44	17 027	64.72	2.56
TDAL	1	**72.80**	0.53	**6 051**	**87.46**	**3.46**
TDAL	1.5	72.41	0.46	10 432	78.39	3.14

increases steeply with the number of manual annotations and reaches a plateau at some point.

Most importantly, the observed TDAL curves are shorter than the DAL ones indicating that the TDAL method requires markedly less manual annotations to achieve the same performance as DAL. In order to demonstrate this cost reduction, we compared the costs in terms of the numbers of manual annotations at the highest UAR achieved by each method (cf. Table 3). Our findings indicate that the proposed TDAL clearly achieved the best performances with 73.69 % UAR, consistently and robustly outperforming the other methods. Comparing our obtained results to those obtained in the INTERSPEECH 2009 Emotion Challenge [52], the TDAL approach outperformed the challenge baseline (67.6 % UAR), as well as the winner of the challenge, who obtained the best result (70.29 % UAR) [61]. Moreover, our novel approach, as well as outperforming these more conventional methods, also reduces manual annotation efforts resulting in the highest annotation cost reduction.

In order to further analyse the obtained results of the proposed algorithms, we calculated, for each algorithm, the average of the maximal UARs over the 20 runs (as given in Table 3), and compared them via a set of a Tukey's post hoc tests[5] to statistically compare the performances (cf. Table 4).

The statistically analysis confirms our previous observations and clearly indicates that our proposed TDAL approach matches with, or significantly outperforms, the other algorithms. Therefore, we can conclude that our proposed algorithm achieves performances tantamount to the other algorithms, with the added advantage of considerably lower annotation costs, thereby saving the associated time and monetary costs.

[5]In this paper, we report basics for conventional Null Hypothesis Testing (NHT), but refrain from a full-fledged NHT analysis due to its inherent problems [62]; instead, we employ effect sizes [63].

TABLE 4. Significance levels for the Tukey's post hoc test obtained for Passive Learning (PL), Dynamic Active Learning (DAL) with agreement level $j = 2$ and $j = 3$, and the trustability-based Dynamic Active Learning (TDAL) with $j = 1$ and $j = 1.5$. Brightest grey indicates effect size $d > 2$, bright grey $d < 2$, dark grey $d < 1$, and black $d < 0$.

[p]	j	PL –	DAL 2	DAL 3	TDAL 1	TDAL 1.5
Active Learning						
PL	–	–				
DAL	2	3.175	–			
DAL	3	5.861	0.202	–		
TDAL	1	4.178	1.713	1.82	–	
TDAL	1.5	4.444	0.142	-0.059	-1.759	–
Upsampled Active Learning						
PL	–	–				
DAL	2	2.683	–			
DAL	3	1.364	-1.521	–		
TDAL	1	4.884	1.12	3.215	–	
TDAL	1.5	2.138	0.293	1.091	-1.383	–
Upsampled Active Learning with Semi-Supervised Learning						
PL	–	–				
DAL	2	0.619	–			
DAL	3	-0.163	-0.649	–		
TDAL	1	1.699	1.096	1.601	–	
TDAL	1.5	0.867	0.281	0.866	-0.786	–

VIII. CONCLUSION AND OUTLOOK

Motivated by a scarcity of annotated data, active learning strategies have been investigated to reduce the cost of gathering labels for databases. In this regard, we introduced the novel Trustability-based Dynamic Active Learning algorithm (TDAL), which incorporates an annotator trustability into a Dynamic Active Learning (DAL) approach. Furthermore, leveraging the advantages of crowdsourcing to collect annotations in a fast and cost-effective manner, we integrated the proposed algorithm into the crowdsourcing platform iHEARu-PLAY [19].

To evaluate our algorithm, we performed emotion recognition studies on the FAU Aibo database [51]. Using a Support-Vector-Machine as the classifier, a passive

learning (PL) approach, acting as a baseline, required more than 48k annotations to achieve a maximum UAR of 60.03 %. In comparison, the DAL approach achieved a relative annotation cost reduction of up to 84.56 % while achieving a UAR of 61.41 %. Moreover, our proposed TDAL saved up to 90.57 % labelling efforts, stopping the annotation process after only 4.5 k collected annotations. Further, the TDAL also outperformed the baseline, achieving a maximum UAR of 62.66 %.

Further experiments were performed making use of upsampling in each AL step in order to obtain an equal class distribution. As a result, the DAL method achieved a lower relative annotation cost reduction of 59.77 %. The TDAL algorithm, however, achieved a greater cost reduction of 78.07 %, while producing a higher maximum classification performance with 73.71 % UAR.

Finally, a Semi-Supervised-Learning step was introduced after each Active Learning step. This approach achieved better classifier performances, but at a smaller cost reduction. Via a combination of upsampling and Semi-Supervised Learning, the DAL algorithm achieved a relative cost reduction of 77.48 %. The TDAL approach was able to reduce the costs to 87.46 %. Again, the TDAL approach outperformed the other algorithms achieving a maximum UAR of 72.80 %, while the DAL and PL achieved 72.29 % UAR and 72.08 % respectively.

The performed experiments indicate that the proposed TDAL algorithm offers clear advantages over the PL method and the conventional DAL approach. While achieving better performances, the main aspect is the effective way of considerably reducing the number of needed annotations and therefore the need for manual labellers, as well as the associate monetary costs. The caveat has to be made that this is a pilot study, conducted on a limited number of datasets and it is applied to the one task of emotion recognition. Therefore, future work will focus on evaluating the TDAL on more databases, as well on evaluating the TDAL approach with even more diverse user trustability scores to demonstrate its robustness and performance improvements. Furthermore, an additional experiment on a task with a continuous label value range can be conducted to investigate the usability of the trustability-weighted median voting method compared to the conventional median.

Summarising the herein presented work, the obtained results lend further weight to the assumption that the TDAL algorithm is an effective approach combining Active Learning and the annotator trustability and can therefore be used in crowdsourcing platforms in order to reduce the annotation costs for emotion recognition tasks while at the same time improving the classification results.

ACKNOWLEDGEMENT

We thank all iHEARu-PLAY users for donating their annotations. Further, we thank Dr. Zixing Zhang at Imperial College London/UK for providing helpful initial baseline code for the active learning algorithms.

REFERENCES

[1] V. Ambati, S. Vogel, and J. Carbonell, "Active learning and crowd-sourcing for machine translation," in *Proc. Int. Conf. Lang. Resour. Eval.*, Valletta, Malta, 2010, pp. 2169–2174.

[2] V. Raykar *et al.*, "Learning from crowds," *J. Mach. Learn. Res.*, vol. 11, pp. 1297–1322, Apr. 2010.

[3] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing for usability: Using micro-task markets for rapid, remote, and low-cost user measurements," in *Proc. Int. Conf. Hum. Factors Comput. Syst.*, Florence, Italy, 2008, pp. 1–4.

[4] A. Tarasov, S. J. Delaney, and C. Cullen, "Using crowdsourcing for labelling emotional speech assets," in *Proc. W3C Workshop Emotion Markup Lang.*, Paris, France, 2010, pp. 1–6.

[5] B. Settles, "Active learning literature survey," Dept. Comput. Sci., Univ. Wisconsin–Madison, Madison, WI, USA, Tech. Rep. 1648, 2009.

[6] Z. Zhang, N. Cummins, and B. Schuller, "Advanced data exploitation in speech analysis: An overview," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 107–129, Jul. 2017.

[7] S. Hantke, E. Marchi, and B. Schuller, "Introducing the weighted trustability evaluator for crowdsourcing exemplified by speaker likability classification," in *Proc. Int. Conf. Lang. Resour. Eval.*, Portoroz, Slovenia, 2016, pp. 2156–2161.

[8] M. D. Smucker and C. P. Jethani, "The crowd vs. the lab: A comparison of crowd-sourced and University Laboratory participant behavior," in *Proc. SIGIR Workshop Crowdsourcing Inf. Retr.*, Beijing, China, 2011, pp. 9–14.

[9] S. Novotney and C. Callison-Burch, "Cheap, fast and good enough: Automatic speech recognition with non-expert transcription," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, Los Angeles, CA, USA, 2010, pp. 207–215.

[10] B. Settles, "From theories to queries: Active learning in practice," in *Proc. Act. Learn. Exp. Design Workshop, Satell. Int. Conf. Artif. Intell. Stat.*, Sardinia, Italy, 2011, pp. 1–18.

[11] L. Yi *et al.*, "A scalable active framework for region annotation in 3D shape collections," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 210:1–210:12, 2016.

[12] Y. Chen *et al.*, "An active learning-enabled annotation system for clinical named entity recognition," *BMC Med. Inform. Decis. Making*, vol. 17, no. 2, p. 82, 2017.

[13] E. Boiy and M.-F. Moens, "A machine learning approach to sentiment analysis in multilingual Web texts," *Inf. Retr.*, vol. 12, no. 5, pp. 526–558, 2009.

[14] Y. Zhang, E. Coutinho, Z. Zhang, C. Quan, and B. Schuller, "Dynamic active learning based on agreement and applied to emotion recognition in spoken interactions," in *Proc. Int. Conf. Multimodal Interact.*, Seattle, WA, USA, 2015, pp. 275–278.

[15] Z. Zhang, F. Eyben, J. Deng, and B. Schuller, "An agreement and sparseness-based learning instance selection and its application to subjective speech phenomena," in *Proc. Workshop Emotion Social Signals, Sentiment Linked Open Data, Satell. Conf. Lang. Resour. Eval.*, Reykjavik, Iceland, 2014, pp. 21–26.

[16] Y. Yan, R. Rosales, G. Fung, and J. G. Dy, "Active learning from crowds," in *Proc. Int. Conf. Mach. Learn.*, Bellevue, WA, USA, 2011, pp. 1161–1168.

[17] F. Laws, C. Scheible, and H. Schütze, "Active learning with amazon mechanical turk," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Edinburgh, U.K., 2011, pp. 1546–1556.

[18] J. Costa, C. Silva, M. Antunes, and B. Ribeiro, "On using crowdsourcing and active learning to improve classification performance," in *Proc. Int. Conf. Intell. Syst. Design Appl.*, Córdoba, Spain, Nov. 2011, pp. 469–474.

[19] S. Hantke, F. Eyben, T. Appel, and B. Schuller, "iHEARu-PLAY: Introducing a game for crowdsourced data collection for affective computing," in *Proc. Int. Workshop Autom. Sentiment Anal. Wild, Satell. Biannu. Conf. Affect. Comput. Intell. Interact.*, Xi'an, China, 2015, pp. 891–897.

[20] S. Hantke, Z. Zhang, and B. Schuller, "Towards intelligent crowdsourcing for audio data annotation: Integrating active learning in the real world," in *Proc. 18th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Stockholm, Sweden, 2017, pp. 3951–3955.

[21] A. K. McCallumzy and K. Nigamy, "Employing EM and pool-based active learning for text classification," in *Proc. Conf. Mach. Learn.*, Madison, WI, USA, 1998, pp. 359–367.

[22] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J. Mach. Learn. Res.*, vol. 2, pp. 45–66, Nov. 2001.

[23] M. Bloodgood and C. Callison-Burch, "Bucking the trend: Large-scale cost-focused active learning for statistical machine translation," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, Uppsala, Sweden, 2010, pp. 854–864.
[24] S. C. Hoi, R. Jin, J. Zhu, and M. R. Lyu, "Batch mode active learning and its application to medical image classification," in *Proc. Int. Conf. Mach. Learn.*, Orlando, FL, USA, 2006, pp. 417–424.
[25] G. Schohn and D. Cohn, "Less is more: Active learning with support vector machines," in *Proc. Int. Conf. Mach. Learn.*, Stanford, CA, USA, 2000, pp. 839–846.
[26] A. Brew, D. Greene, and P. Cunningham, "Using crowdsourcing and active learning to track sentiment in online media," in *Proc. Conf. Artif. Intell.*, Lisbon, Portugal, 2010, pp. 145–150.
[27] M. Wang and X.-S. Hua, "Active learning in multimedia annotation and retrieval: A survey," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 2, p. 10, 2011.
[28] G. Riccardi and D. Hakkani-Tur, "Active learning: Theory and applications to automatic speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 504–511, Jul. 2005.
[29] Z. Zhang and B. Schuller, "Active learning by sparse instance tracking and classifier confidence in acoustic emotion recognition," in *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Portland, OR, USA, 2012, pp. 1–4.
[30] Y. Zhang, Y. Zhou, J. Shen, and B. Schuller, "Semi-autonomous data enrichment based on cross-task labelling of missing targets for holistic speech analysis," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Shanghai, China, Mar. 2016, pp. 6090–6094.
[31] Y. Yan, R. Rosales, G. Fung, F. Farooq, B. Rao, and J. G. Dy, "Active learning from multiple knowledge sources," in *Proc. Int. Conf. Artif. Intell. Stat.*, La Palma, Spain, 2012, pp. 1350–1357.
[32] D. R. Karger, S. Oh, and D. Shah, "Iterative learning for reliable crowdsourcing systems," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, Granada, Spain, 2011, pp. 1953–1961.
[33] D. E. Difallah, G. Demartini, and P. Cudré-Mauroux, "Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms," in *Proc. Int. Workshop Crowdsourcing Web Search*, Lyon, France, 2012, pp. 26–30.
[34] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Cooperative learning and its application to emotion recognition from speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 1, pp. 115–126, Jan. 2015.
[35] B. Frénay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 5, pp. 845–869, May 2014.
[36] B. Biggio, B. Nelson, and P. Laskov, "Support vector machines under adversarial label noise," in *Proc. Asian Conf. Mach. Learn.*, Taoyuan, Taiwan, 2011, pp. 97–112.
[37] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari, "Learning with noisy labels," in *Advances in Neural Information Processing Systems*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2013, pp. 1196–1204.
[38] V. S. Sheng, F. Provost, and P. G. Ipeirotis, "Get another label? Improving data quality and data mining using multiple, noisy labelers," in *Proc. Int. Conf. Knowl. Discovery Data Mining (SIGKDD)*, Las Vegas, NV, USA, 2008, pp. 614–622.
[39] P. G. Ipeirotis and P. K. Paritosh, "Managing crowdsourced human computation: A tutorial," in *Proc. Int. Conf. Companion World Wide Web*, Hyderabad, India, 2011, pp. 287–288.
[40] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Adv. Large Margin Classifiers*, vol. 10, no. 3, pp. 61–74, 1999.
[41] L. V. Ahn, "Games with a purpose," *Computer*, vol. 39, no. 6, pp. 92–94, Jun. 2006.
[42] F. Provost, "Machine learning from imbalanced data sets 101," in *Proc. AAAI Workshop Learn. Imbalanced Data Sets*, Austin, TX, USA, 2000, pp. 1–3.
[43] Z. Zhang, "Semi-autonomous data enrichment and optimisation for intelligent speech analysis," Ph.D. dissertation, Dept. Mensch-Maschine-Kommuniktion, Tech. Univ. Munich, Munich, Germany, 2015.
[44] C. Sun, N. Rampalli, F. Yang, and A. Doan, "Chimera: Large-scale classification using machine learning, rules, and crowdsourcing," *Proc. VLDB Endowment*, vol. 7, no. 13, pp. 1529–1540, 2014.
[45] G. Demartini, D. E. Difallah, and P. Cudré-Mauroux, "ZenCrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking," in *Proc. Int. Conf. World Wide Web*, Lyon, France, 2012, pp. 469–478.
[46] J. Whitehill, T.-F. Wu, J. Bergsma, J. R. Movellan, and P. L. Ruvolo, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2009, pp. 2035–2043.
[47] P. Welinder, S. Branson, P. Perona, and S. J. Belongie, "The multidimensional wisdom of crowds," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2010, pp. 2424–2432.
[48] E. Kamar, S. Hacker, and E. Horvitz, "Combining human and machine intelligence in large-scale crowdsourcing," in *Proc. Int. Conf. Auton. Agents Multiagent Syst.*, Valencia, Spain, 2012, pp. 467–474.
[49] S. Burke, "Missing values, outliers, robust statistics & non-parametric methods," *Stat. Data Anal.*, vol. 59, no. 1, pp. 19–24, 2001.
[50] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "From game design elements to gamefulness: Defining gamification," in *Proc. Int. Conf. Acad. MindTrek*, Tampere, Finland, 2011, pp. 9–15.
[51] S. Steidl, "Automatic classification of emotion related user states in spontaneous children's speech," Ph.D. dissertation, Dept. Pattern Recognit., Lab., Univ. Erlangen-Nuremberg, Erlangen, Germany, 2009.
[52] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proc. 10th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Brighton, U.K., 2009, pp. 312–315.
[53] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proc. Int. Conf. Multimedia*, Barcelona, Spain, 2013, pp. 835–838.
[54] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Commun.*, vol. 53, pp. 1162–1171, Nov. 2011.
[55] J. Deng, Z. Zhang, and B. Schuller, "Linked source and target domain subspace feature transfer learning–exemplified by speech emotion recognition," in *Proc. Int. Conf. Pattern Recognit.*, Stockholm, Sweden, Aug. 2014, pp. 761–766.
[56] A. Baird *et al.*, "Automatic classification of autistic child vocalisations: A novel database and results," in *Proc. 18th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Stockholm, Sweden, 2017, pp. 849–853.
[57] V. Vapnik, *The Nature of Statistical Learning Theory*. Berlin, Germany: Springer, 2013.
[58] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.
[59] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explorations Newslett.*, vol. 11, no. 1, pp. 10–18, 2009.
[60] B. Schuller, *Intelligent Audio Analysis*. Berlin, Germany: Springer, 2013.
[61] P. Dumouchel, N. Dehak, Y. Attabi, R. Dehak, and N. Boufaden, "Cepstral and long-term features for emotion recognition," in *Proc. 10th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Brighton, U.K., 2009, pp. 344–347.
[62] R. L. Wasserstein and N. A. Lazar, "The ASA's statement on p-values: Context, process, and purpose," *Amer. Stat.*, vol. 70, no. 2, pp. 129–133, 2016.
[63] G. M. Sullivan and R. Feinn, "Using effect size—Or why the *P* value is not enough," *J. Graduate Med. Educ.*, vol. 4, pp. 279–282, Sep. 2012.

SIMONE HANTKE received the Diploma degree in media technology from the Technische Hochschule Deggendorf in 2011 and the M.Sc. degree from the Technische Universität München, one of Germany's Excellence Universities, in 2014, where she is currently pursuing the Ph.D. degree. She is with the ZD.B Chair of the Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany. She is involved on her Ph.D. thesis in the field of affective computing and speech processing, focusing her research on data collection and new machine learning approaches for robust automatic speech recognition and speaker characterization. Her main area of involvement has been with the EU FP7 ERC Project iHEARu. In the scope of this project, she leads the development of crowdsourcing data collection and annotation for speech processing and is the lead author of iHEARu-PLAY.

ALEXANDER ABSTREITER received the bachelor's degree in computer science from the University of Passau, Germany, in 2017. He is currently pursuing the master's degree in computer engineering with the Politecnico di Torino, Italy, with a focus on data science. His interests are machine learning methods which help to deal with unlabeled data.

NICHOLAS CUMMINS received the bachelor's degree (Hons.) and the Ph.D. degree in electrical engineering from UNSW, Australia, in 2011 and 2016, respectively. He is currently pursuing the Habilitation degree with the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, where he is involved in Horizon 2020 projects DE-ENIGMA, RADAR-CNS, and TAPAS. His current research interests include the areas of behavioral signal processing with a focus on the automatic multisensory analysis and understanding of different health states. During his Ph.D. degree, he investigated whether the voice can be used as an objective marker in the diagnosis and monitoring of clinical depression. He has published regularly in the field of depression detection since 2011; these papers have attracted significant attention and citations.

BJÖRN SCHULLER received the Diploma degree in 1999, the Ph.D. degree in automatic speech and emotion recognition in 2006, and the Habilitation degree and the Adjunct Teaching Professorship in signal processing and machine intelligence in 2012, all in electrical engineering and information technology from TUM, Munich, Germany. He is a Reader in machine learning with the Department of Computing, Imperial College London, U.K., a Full Professor and the Head of the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany, and a CEO of audEERING–an Audio Intelligence company. He has (co-)authored six books and over 650 publications in peer-reviewed books, journals, and conference proceedings leading to more than overall 16 000 citations (h-index = 63). He is a President-Emeritus of the AAAC, and an elected member of the IEEE Speech and Language Processing Technical Committee. He is a Co-Program Chair of Interspeech 2019, the Area Chair of ICASSP, and the Editor-in-Chief of the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING next to a multitude of further an associate and the guest editor roles and functions in technical and organizational committees.