LET'S AGREE TO DISAGREE: CONSENSUS ENTROPY ACTIVE LEARNING FOR PERSONALIZED MUSIC EMOTION RECOGNITION

Juan Sebastián Gómez-Cañón[1] **Estefanía Cano** [2] **Yi-Hsuan Yang** [3]
Perfecto Herrera [1] **Emilia Gómez** [4,1]
[1] Music Technology Group, Universitat Pompeu Fabra, Spain
[2] Songquito UG, Erlangen, Germany
[3] Academia Sinica, Taiwan
[4] Joint Research Centre, European Commission, Seville, Spain
juansebastian.gomez@upf.edu

ABSTRACT

Previous research in music emotion recognition (MER) has tackled the inherent problem of subjectivity through the use of personalized models – models which predict the emotions that a particular user would perceive from music. Personalized models are trained in a supervised manner, and are tested exclusively with the annotations provided by a specific user. While past research has focused on model adaptation or reducing the amount of annotations required from a given user, we propose a methodology based on uncertainty sampling and query-by-committee, adopting prior knowledge from the agreement of human annotations as an oracle for active learning (AL). We assume that our disagreements define our personal opinions and should be considered for personalization. We use the DEAM dataset, the current benchmark dataset for MER, to pre-train our models. We then use the AMG1608 dataset, the largest MER dataset containing multiple annotations per musical excerpt, to re-train diverse machine learning models using AL and evaluate personalization. Our results suggest that our methodology can be beneficial to produce personalized classification models that exhibit different results depending on the algorithms' complexity.

1. INTRODUCTION

Historically, the field of MER has mainly focused on extracting meaningful acoustic features from audio and associating them to possible emotions that music can convey [1]. Machine learning algorithms are trained with these features and then linked with the emotional judgements that annotators report to perceive or feel when listening to the music [2] – ultimately presented as "ground truth" to the algorithms. One of the key issues to MER is the difference between perceived and induced emotions: perceived emotions are the listeners' judgements with respect to musical properties (e.g., key, tempo, timbre), while induced (or felt) emotions are those that the music may arouse within the listener. Despite the effort from the field of music cognition to better understand the psychological differences between these emotions [3–6], their contrast poses a fundamental obstacle to the field of MER. Namely, the construction of the needed "ground truth" results questionable: (1) listeners are commonly confused between perceived and induced emotions when reporting their emotional judgements, (2) the inherent subjectivity from the annotation task is typically addressed by averaging the several annotations into a common "ground truth", and (3) the annotation procedure is a highly demanding task resulting in small datasets with few annotations per music excerpt. Nonetheless, researchers from the field of MER have tackled this problem by: (1) identifying whether the listener's response is based on the judgement of perceived or induced emotions [7] and attempting to train listeners [8], (2) using exclusively the emotion reports from a particular listener or group of listeners to produce personalized and group-based models [9], and (3) introducing AL methods to reduce the amount of annotations required to train such algorithms (see Section 2). Given the importance the construction of a "ground truth" for MER, we address two research questions in this paper:

RQ1 - Can we exploit human agreement in music emotion annotations as input for AL methodologies to produce personalized models?

RQ2 - What is the impact of the choice of classification algorithm on personalization of MER systems?

The rest of this paper is structured as follows: Section 2 reviews basic definitions and previous work, in Section 3 we detail the methodology of our study, including the proposed consensus entropy methods and classification schemes. Section 4 provides results of our study which are later discussed in Section 5.

2. RELATED WORK

Individual differences of listeners have a significant impact on the performance of a MER algorithm. To this

© J.S. Gómez-Cañón, E. Cano, Y.-H. Yang, P. Herrera and E. Gómez. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** J.S. Gómez-Cañón, E. Cano, Y.-H. Yang, P. Herrera and E. Gómez, "Let's agree to disagree: Consensus Entropy Active Learning for Personalized Music Emotion Recognition", in *Proc. of the 22nd Int. Society for Music Information Retrieval Conf.,* Online, 2021.

extent researchers have proposed two distinct solutions to tackle the subjectivity issue [1,9]: group-based and personalized MER models. Group-based MER assembles annotators according to individual factors (e.g., sex, age, music experience) and create a common "ground truth" for this group. Personalized MER uses the annotations from a specific user to train a machine learning model. Yang et al. [9] tested both approaches for the regression task and found that: (1) group-based methods do not outperform general models (i.e., models which are trained with the common "ground truth" from the complete set of users), and (2) personalized algorithms largely outperform general models. However, Gómez-Cañón et al. [10] studied the influence of native language and self-reported lyrics comprehension on the agreement of annotations and their impact on group-based MER classification. The authors found substantial differences in the annotations of users with different mother tongues (consistent with findings from [11, 12]), and a direct impact of individual differences (i.e., familiarity, preference, and lyrics comprehension) on the agreement of these annotations. The authors also reported that group-based MER algorithms trained on the annotations of users that reported understanding the lyrics, consistently outperformed general models for a small dataset with a large amount of annotations per excerpt, contradicting results by Yang et al. [9]. More research is needed on the topic of group-based MER, hence in this paper we focus on the need of personalization strategies.

Su and Fung [13] proposed using AL (i.e., uncertainty sampling) in order to achieve personalization – which is the focus of this paper. The aim of AL is to minimize the annotation cost by cleverly choosing unlabeled data instances, such that machine learning algorithms perform better with less training [14]. Sarasúa et al. [15] used it to reduce the amount of training instances and achieve better classification performance for MER. Uncertainty sampling uses the posterior probability from a classification model to assess the most difficult/uncertain unlabeled data instances (e.g., consider an output probability of 0.5 for binary classification). [1] Su and Fung [13] used two sampling methods to select training instances: (1) using the most *informative* instances – with highest uncertainty, and (2) using the most *representative* instances – with least uncertainty. Their results suggested that AL can reduce the annotation task up to 80% without decreasing performance of classification. However, the performance of AL as a personalization strategy appears to be hindered by low quality annotations – a problem known as the "noisy oracle" issue: low reliability in annotations results in poor training instances, in turn resulting in poor classification performance. In this direction, multi-oracle AL [17, 20–23] has been proposed to exploit multiple annotators by estimating the importance of both unlabeled instances and the expertise of each annotator – ultimately improving label quality. More recently, Chen et al. [24, 25] proposed model adaptation to achieve personalization. Their approach relied on developing a general MER regression model (namely, Gaussian Mixture Models) and progressively tying the Gaussian components to adapt the models based on the maximum a posteriori (MAP) linear regression. Results evidence that only 10-20 personal annotations are necessary to obtain the same level of accuracy as a baseline model (50 annotations). However, they found no statistically significant difference between the proposed tying methods. Overall, we find two limitations in the MER personalization literature: (1) the evaluation of different AL strategies and (2) the definition of best algorithms for effective personalization.

[1] We refer the reader to [14, 16–19] for a comprehensive overview of AL methods.

3. METHODOLOGY

The main contribution of our work is to address open questions by proposing query strategies that involve collective judgement for personalization and evaluating diverse algorithms, later introduced in Section 3.3. We use a different query strategy to build upon the work by Su and Fung [13], and propose a novel method to account for the collective judgement – differing from traditional instance selection for AL. Our work is also motivated by the multi-oracle AL paradigm [20–22] in order to exploit this judgement. However, instead of picking an expert/confident annotator, we select instances which are ambiguous to the crowd – different to those ambiguous to the algorithms. We introduce *consensus entropy* [26] to AL for MER with a threefold perspective: (1) analyzing the agreement achieved by a committee of pre-trained models (*machine consensus - MC*), (2) analyzing the agreement from a committee of annotators (*human consensus - HC*), and (3) taking into account both committees (*hybrid consensus - MIX*). Our work differs from [25] since we obtain personalization by sampling informative instances and re-training the algorithms, instead of progressively adapting model parameters. Our main assumption is that prior knowledge about the uncertainty of an excerpt with respect to the collective judgement (i.e., human consensus), results in the particular instances which could be indicative of classification boundaries across individual listeners. Music excerpts on which we disagree upon define our personal opinions and should be taken into account for personalization. Secondly, we assume that the confusion between perceived and induced emotions is mainly static and will not vary over time (see [27] for a study on intra-rater agreement), hence personalization could lead to models that can predict both types of emotion and work must be done to determine the type of emotion [7, 8]. To the best of our knowledge, the use of the collective judgement as a personalization strategy has never been explored in MER so far. [2]

3.1 Data

Despite the complexity and difficulty of obtaining music emotion annotations, researchers in MER have made great efforts to create open datasets. [3] To pre-train our classifiers, we used the DEAM dataset [28]. The benchmark

[2] https://github.com/juansgomez87/consensus-entropy

[3] Data from the study in [13] is not openly available.

dataset for MER, DEAM was constructed across several MediaEval contests (2013–2015), and contains 1802 music excerpts and dynamic arousal and valence annotations (introduced by Russell [29]). We discretized annotations into four quadrants for classification, following [30]: Q1 (positive valence and arousal, A+V+), Q2 (positive arousal and negative valence, A+V-), Q3 (negative valence and arousal, A-V-), Q4 (negative arousal and positive valence, A-V+). [4] To test personalization, we used the AMG1608 dataset [35]. This dataset was previously used for personalization purposes [24,25], and is composed of 1608 music excerpts rated with static arousal-valence annotations from 665 listeners (22 annotators from the campus of the National Taiwan University and 643 from Amazon Mechanical Turk). From the pool of annotators, we use the subset of 46 annotators that rated more than 150 songs (from which 10 belong to the campus subset).

We used two feature sets depending on the classification algorithm (see Section 3.3): (1) low-level, emotionally-relevant features for classic machine learning algorithms, and (2) mel-spectrograms for novel convolutional neural network architectures. As to (1), the IS13 ComParE feature set [36] has been widely used for sound, speech, and music emotion recognition. We extracted 260 features (mean and standard deviation of 65 low-level music descriptors and their first order derivatives) from segments of 1 second [28], with 50% overlap, and standardize across features – using OpenSMILE [37]. In order to test our approach on novel deep learning architectures we extracted mel-spectrograms, based on [38]: we downsampled audio to 16kHz, performed a Short-Time Fourier Transform (window size: 512 samples $\sim$ 23ms; hop size: 256 $\sim$ 12ms), and extracted a mel-scale spectrogram with 128 mel-bands – using Librosa [39].

3.2 Consensus entropy

Consensus entropy is a combination of uncertainty sampling and query-by-committee methods as follows [16,26]: (1) a committee of classifiers predicts the output probabilities of unlabeled data, (2) probabilities are averaged across the committee of classifiers, (3) uncertainty is calculated as Shannon's entropy across classes for each instance, (4) q instances with highest entropy are selected to be annotated by the oracle, and (5) classifiers are retrained with the provided annotations. For example, full disagreement from a committee of four classifiers results when each one predicts a different quadrant with 100% probability. This yields average probabilities per quadrant $p_{avg} = \{Q1 : 0.25, Q2 : 0.25, Q3 : 0.25, Q4 : 0.25\}$ and high inter-class entropy/uncertainty of 1.386. We refer to this approach as *machine consensus (MC)*. Secondly, studies have shown evidence of the impact of inter-rater agreement on the performance of MER algorithms [10,12]. Hence, we propose *human consensus (HC)* as a variation from classical consensus entropy: we calculate entropy on the normalized annotation histogram per song. For example, given 6 annotators for song i, we obtain a relative frequency $f_i = \{Q1 : 1/6, Q2 : 2/6, Q3 : 3/6, Q4 : 0/6\}$. Thirdly, we combine the strategies for a *hybrid consensus (MIX)* by stacking the probabilities and relative frequencies, and calculating the overall entropy.

The proposed method is summarized in Algorithm 1: let $\mathcal{L} = \left\{(x_i, y_i)\right\}_{i=1}^{m}$ represent the pre-training data (DEAM) consisting of m labeled instances (1802 excerpts) and $\mathcal{U} = \left\{(x_i)\right\}_{i=m+1}^{n}$ represent the "unlabeled" data for personalization (AGM1608). Since $\left\{(y_i)\right\}_{i=m+1}^{n}$ for $\mathcal{U}$ are already present in the dataset, we query q excerpts and fine-tune with their annotations. We consider x_i an input feature (low-level feature vector or mel-spectrogram), and $y_i \in C = \{Q1, Q2, Q3, Q4\}$ as the annotated quadrant. Finally, we split annotated data by each user $u_j = \{0 \dots 45\}$ with more than 150 annotations: 85% for training and 15% for testing, with no overlapping music excerpts. We denote $P_{\mathcal{L},M_k}(y_i|x_i)$ as the conditional probability of y given x according to a classifier M_k trained on $\mathcal{L}$. Notice that $P_{\mathcal{L}}(y_i|x_i)$ is the probability averaged across all models M_k. We performed 10 iterations and queried $q = 10$ instances per iteration. [5] Following Chen et al. [25], we used random selection as a baseline.

Algorithm 1: Consensus entropy for MER.

```
input : Labeled data L, unlabeled data U
Pre-train each model M_k on L;
for each iteration it = {0...9} do
    for each user u_j do
        Calculate P_{L,M_k}(y_i|x_i) for each x_i ∈ U;
        if MC then
            Average P_L(y_i|x_i) across frames and
              M_k models;
            Select q excerpts with highest entropy;
        else if HC then
            Calculate relative frequency f_i per
              music excerpt;
            Select q excerpts with highest entropy;
        else if MIX then
            Calculate and stack P_L(y_i|x_i) and f_i;
            Select q excerpts with highest entropy;
        else
            Select q random excerpts;
        Annotate q instances by u_j;
        for each model M_k do
            for each (x_i, y_i) ∈ q do
                Re-train M_k on L ∪ (x_i, y_i);
                Compute metrics on test data;
                Update L ← L ∪ (x_i, y_i) and
                  U ← U \ (x_i, y_i);
            end
        end
    end
end
```

[4] We refer the reader to [10] for a concise explanation of music emotion taxonomies. See also [31–34] for in-depth theory.

[5] Tests using $q = 15$ and $q = 40$ reduced the amount of available users (i.e., each user annotated a different amount of excerpts). We chose $q = 10$ to match the study in [25]: 46 users with over 150 annotations.

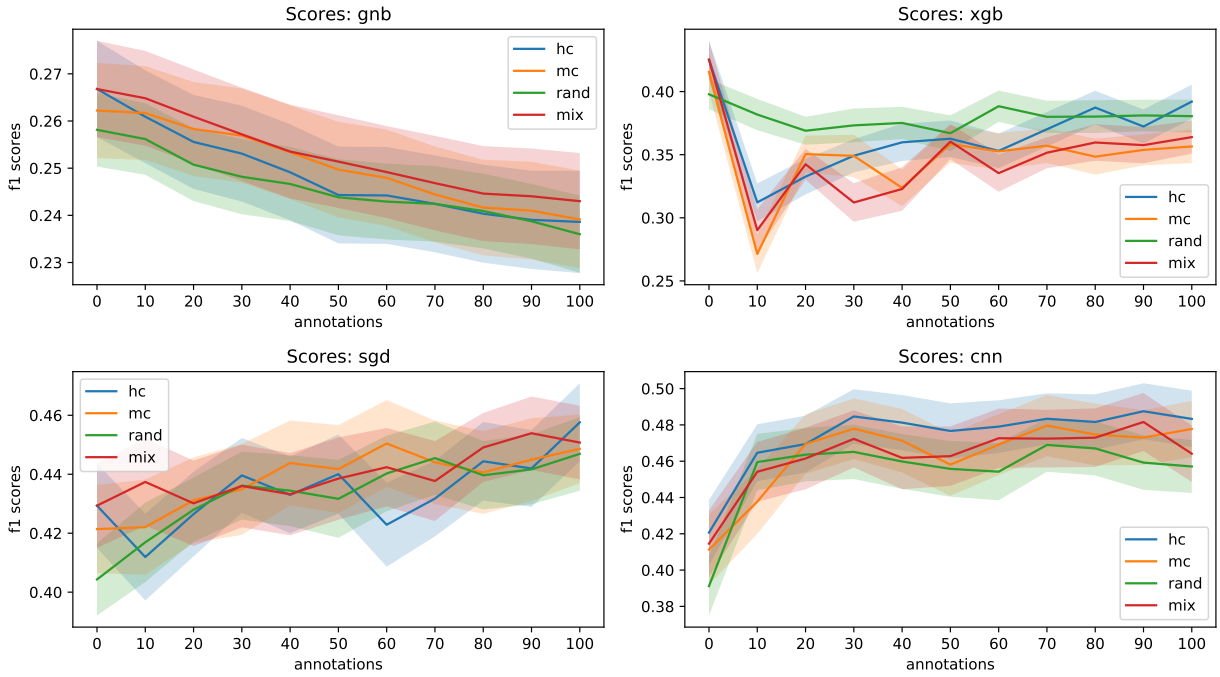

Figure 1. Average results of weight-averaged F1-scores for each type of model, across 46 users and 5 classifiers (shaded area corresponds to $CI = 95\%, n = 230$). HC stands for Human Consensus, MC for machine consensus, MIX for hybrid consensus and RAND for random selection.

3.3 Algorithms

Since the query strategy requires a committee of classifiers, we pre-trained all the following models with the DEAM dataset using 5-fold cross validation – each classifier is pre-trained on general annotations while still resulting in diverse predictions, in order to analyze agreement amongst classifiers. For each algorithm $m_k = \{0 \dots 4\}$, we obtained 5 classifiers for a total of 20 models per user. We used four algorithms in this study – based on (1) computational efficiency and low memory cost, and (2) well-established and novel approaches in the state-of-the-art – and introduce them as follows:

Gaussian Naive Bayes (GNB). These algorithms are based on the "naive" assumption of independence between pairs of features, given a class label [40]. Bayes' theorem relates the conditional probability of the output y and the dependent feature vectors x_i. The likelihood of the features is assumed to be normal-distributed, hence the models are Gaussian. Priors are adjusted according to the data and variance smoothing is set to 1e-9.

Extreme Gradient Boosting (XGB). This widely used machine learning method is based on the idea of gradient tree boosting: an ensemble of weak learners (i.e., regression trees) is optimized to minimize a given loss function [41]. In contrast to other gradient boosting algorithms, XGB is well-established given its scalability and training speed. We performed a minor change to allow re-fitting the algorithm and set parameters empirically during pre-training: the maximum depth of 5 for each decision tree.

Logistic Regression (SGD). We used a model that optimizes a log-loss function with L2 regularization to output class probabilities – obtaining a Logistic Regression classifier fitted using Stochastic Gradient Descent (SGD). SGD is an optimization method to fit linear classifiers using convex loss functions [42].

Short-chunk Convolutional Neural Network (CNN). In the field of automatic audio tagging, Won et al. [38] have recently proposed a 7-layer 2D convolutional neural network that processes chunks of 3.69s of audio and (2×2) max-pooling layers to summarize the chunk into a single dimension. A mixture of scheduled Adam [43] and SGD are used as optimization methods, following [44]. We pre-trained models for 200 epochs and re-trained for 100 epochs – best models were selected when the validation loss improved.

4. RESULTS

Figure 1 shows the weighted-average F1-scores on the test data averaged across 46 users, averaging across each algorithm for a total of 3680 trained classifiers (46 users $\times$ 4 algorithms $\times$ 5 models per pre-training split $\times$ 4 consensus entropy methods). We report weighted average scores since datasets are class-imbalanced.

4.1 Algorithms and consensus entropy methods

Firstly, we use pairwise, one-sided t-tests ($d.f. = 229$, statistical significance $p < 0.05$) in order to evaluate differences among consensus entropy methods (i.e., HC, MC, MIX, and RAND) for each particular model after 100 annotations (at least 150 annotations are available per user), as evaluated by Chen et al. [25]. We do not perform other

statistical tests (i.e., McNemar's Test or Wilcoxon signed-rank test), as proposed by Demšar [45], since each user has annotated different songs (i.e., the training and testing data is not the same between users).

Gaussian Naive Bayes (GNB). These classifiers appear to diminish their performance with more annotations which is expected of naive bayesian models (i.e., limited generalization to new data) – MIX appears to outperform the random baseline by ~ 1 percent point. However, none of the comparisons between methods is statistically significant ($p > 0.147$).

Extreme Gradient Boosting (XGB). These classifiers display an expected behavior: random selection results in limited variation throughout 100 annotations, while other methods (MC, HC, and MIX) suffer a significant fall with the initial re-training data and increasingly improve with the amount of annotations. In this case, HC is significantly better than MC ($p = 0.0001$) and than MIX ($p = 0.0017$), but does not outperform RAND.

Logistic Regression (SGD). These classifiers exhibit increasing performance with more data – the HC method outperforms the random baseline by ~ 1 percent points. Again, none of the pairwise comparisons show significant differences across cross entropy methods ($p > 0.125$).

Short-chunk Convolutional Neural Network (CNN). Classifiers exhibit a significant increase with initial re-training data – the HC method again outperforms the random baseline by ~ 2 percent points. Interestingly, these classifiers display the best performance across all models with cases of high f1-scores (approximately 0.7-0.8 for particular users – see Figure 2). HC is significantly better than RAND ($p = 0.00811$) and than MIX ($p = 0.044$), while MC is better than RAND ($p = 0.0291$). Similar to results reported by Chen et al. [25], these classifiers appear to improve after 20-30 annotations and plateau.

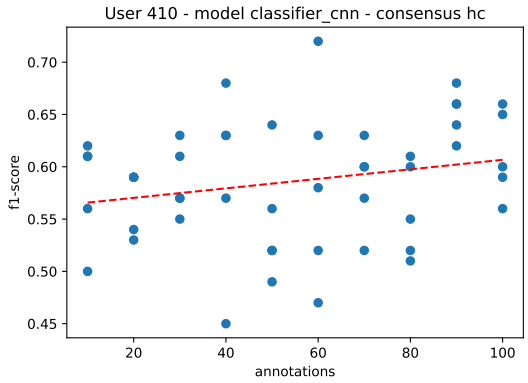

Figure 2. F1-scores of five CNN classifiers from user 410 using the HC consensus entropy method. Each point represents the F1-score for each classifier on the user's test data.

4.2 Campus subset

Given the impact of agreement on classification performance, Chen et al. [25] split the users into two groups: the general pool of 46 annotators and subset of 10 annotators from campus (as mentioned in Section 3.1). We perform the same analysis for this subset of annotators ($d.f. = 49$, statistical significance $p < 0.05$). With respect to the subset of campus annotators, the general tendencies mentioned in Section 4.1 appear to hold, yet the difference of performance between the proposed methods and the random baseline appears to narrow. [6] For the XGB model, HC significantly outperforms MIX ($p = 0.0149$). For the CNN model, HC significantly outperforms RAND ($p = 0.0254$) and MIX ($p = 0.0152$).

4.3 Effective personalization

Although the proposed methods marginally outperform the random selection baseline in the general behavior across all users, we observe diverse behaviors when analyzing each user: (1) XGB and SGD classifiers exhibit less variation across each algorithm than GNB and CNN – XGB and SGD classifiers appear to be more stable with respect to each re-training iteration, and (2) models do not necessarily improve with more annotations – it is likely that the annotations of a particular user are not producing personalization. [7] Thus, we tested evaluating each user's algorithms as seen in Figure 2 and fitted a linear regression (using Ordinary Least Squares) to estimate if the average metrics from the ensemble of classifiers indeed improved as more personal annotations are presented. Namely, when the slope of the lineal regressor is positive, we assume that "effective" personalization has been achieved – as more personal annotations are presented, the algorithm improves performance on the test data. Figure 3 summarizes the results of the amount of personalized models following this assumption: (1) GNB classifiers are rarely producing personalized models, (2) SGD classifiers appear to produce the same number of personalized models regardless the consensus entropy method (slightly more personalization is achieved with HC), and (3) for both XGB and CNN classifiers all proposed consensus entropy methods appear to produce more personalized models than RAND.

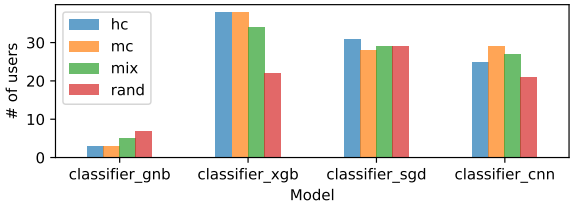

Figure 3. Number of users with effective personalization per algorithm from a total of 46 users.

5. DISCUSSION AND CONCLUSIONS

5.1 Discussion

Our study is inspired by the work from Su and Fung [13], in which AL was used to progressively re-train MER mod-

[6] Refer to Figure 1 from supplementary material.
[7] Refer to Figures 2-4 from supplementary material.

els with informative and representative data instances to produce such personalized models. We also are encouraged by studies from Bullard et al. [46] in which the traditional approach of AL is challenged in favor of "realistic human interaction": instead of providing the algorithms an *optimal query strategy* [14], we attempt to *put the human-in-the-loop* by grounding emotion concepts based on community judgements and simple inter-rater agreement. We aimed at using the collective judgement of the pool of annotators as prior knowledge for AL – our main assumption is that highly uncertain instances in the collective judgement reflect individual boundaries of classification that should be used to personalize MER models. Thus, we propose two consensus entropy methods for AL based on the classical uncertainty sampling and query-by-committee strategies: (1) *human consensus (HC)* that uses the pool of annotators as the committee to obtain informative samples, and (2) *hybrid consensus (MIX)* that considers possible complementary advantages from HC and MC.

Regarding *RQ1* - Can we exploit human agreement in music emotion annotations as input to AL methodologies to produce personalized models? Our findings suggest that our proposed methods appear to improve personalization with respect to a baseline that presents random instances for re-training. Particularly, the proposed HC method outperforms the methods presented by Su and Fung [13], which rely on using uncertainty sampling to compare most informative (highest entropy - MC) data instances for personalization for 8 users. Their study reports average F1-scores of $\mu = 0.35, \sigma = 0.30$) after using AL. Our study shows the following F1-scores from 46 users: **CNN** – $\mu = 0.48, \sigma = 0.12$, **XGB** – $\mu = 0.39, \sigma = 0.10$, **SGD** – $\mu = 0.457, \sigma = 0.10$, GNB – $\mu = 0.238, \sigma = 0.08$. Additionally, the MIX method marginally outperforms the random baseline, showing similar performance to the MC method – the MIX method is likely querying similar instances as the MC method for each iteration.

With respect to *RQ2* - What is the impact of the choice of classification algorithm on personalization of MER systems? We tested our method on four types of algorithms, which display different behaviours: (1) Gaussian Naive Bayes classifiers (GNB) appear not to generalize to new data or work for personalization – Naive Bayes assumes independence of predictors which is not likely the case for overlapping emotionally-relevant features, (2) Logistic Regression (SGD) appears to produce personalized models but there is no significant difference across the consensus entropy methods – the assumption of linearity between features and annotations is not likely to capture more complex relationships from features, (3) Extreme Gradient Boosting classifiers (XGB) appear to produce the highest amount of personalized models – however, results suggest that these models require more annotations in order to eventually surpass the performance metrics from the random selection baseline, and (4) Short-chunk Convolutional Neural Network (CNN) appears to produce the best classification performance and the HC method appears to produce more personalized models – yet the "black box" nature of neural networks might hinder the interpretability and explainability of using these models.

In addition to the fact that our findings are limited by the datasets and the methodologies used to build and annotate them, we present three main limitations to be considered:

Inter-rater agreement. Previous studies [10,11,27,47] have evaluated inter-rater agreement as defined by Krippendorff's coefficient α [48]. However, it is not possible to use this coefficient to assess agreement for the HC method since the annotations are categorical. Only one coefficient can be calculated for arousal or valence over the complete dataset (or dataset subset). Nonetheless, the relative frequency (HC) can be interpreted as an empirical probability that is informative with respect to simple agreement.

Interpretability of the MC method. The lack of agreement between the classifiers might be due to other factors different than the difficulty of the "ground truth". In this sense, acoustic properties and the impact of the features on predictions might produce confounding factors for the classifiers and will be considered as future work.

Stasis of the HC method. HC is mainly static as opposed to MC approach: the method is restricted by the amount of annotated songs and number of users. Thus, the songs that result from each query will be the same for all users, as opposed to the MC approach. In the case of MC, every time a model is re-trained the classification boundaries are adjusted along with the uncertainty of new particular instances. Although the underlying principles of HC and MC are quite different, the expectation of complementary advantages over each other was not met.

5.2 Conclusions

To the extent of our knowledge, the proposed methodology has not been used for the MER task or other MIR use cases, since the classic aim of AL is to make the data collection less burdensome (i.e., reduce the workload of the annotation procedure). In the context of producing user-centric MIR [49], we argue that using knowledge about the collective consensus could be beneficial for other tasks with low inter-rater agreement: music auto-tagging [50], music similarity [11, 27], automatic chord estimation [51], and beat tracking [52]. Indeed, current streaming services and social media constantly produce high amounts of diverse responses in tagging environments – which could be beneficial to test our methodology on other tasks. For the particular field of MER, building a collective "ground truth" by merely averaging ratings across annotators might be oversimplifying what has recently been questioned in neuroscience research on emotions by Barrett [53]:

> "One instance of [an emotion] need not look or feel like another, nor will it be caused by the same neurons [in the brain]. *Variation is the norm.* Your range of [an emotion] is not necessarily the same as mine, although if we were raised in similar circumstances, we will likely have some overlap."

For once, we could simply agree to disagree.

6. ACKNOWLEDGEMENTS

The research work conducted in the Music Technology Group at the Universitat Pompeu Fabra is partially supported by the European Commission under the TROMPA project (H2020 770376). We would like to thank anonymous reviewers that helped us improve the paper with constructive feedback.

7. REFERENCES

[1] Y.-H. Yang and H. H. Chen, *Music Emotion Recognition*. CRC Press, 2011.

[2] N. N. Vempala and F. A. Russo, "Modeling music emotion judgments using machine learning methods," *Frontiers in Psychology*, vol. 8, 2018.

[3] A. Gabrielsson, "Emotion perceived and emotion felt: Same and different," *Musicae Scientiae*, vol. 10, no. 2, pp. 191–213, 2006.

[4] J. Cespedes-Guevara and T. Eerola, "Music communicates affects, not basic emotions - A constructionist account of attribution of emotional meanings to music," *Frontiers in Psychology*, vol. 9, no. FEB, pp. 1–19, 2018.

[5] L. A. Warrenburg, "Comparing musical and psychological emotion theories," *Psychomusicology: Music, Mind, and Brain*, vol. 30, no. 1, pp. 1–19, 2020.

[6] ——, "Choosing the right tune: A review of music stimuli used in emotion research," *Music Perception*, vol. 37, no. 3, pp. 240–258, 2020.

[7] J. S. Gómez-Cañón, N. Gutiérrez-Páez, L. Porcaro, A. Gkiokas, P. Herrera, and E. Gómez, "Improving emotion annotation of music using citizen science," in *Proceedings of the 16th International Conference on Music Perception and Cognition (ICMPC/ESCOM)*, Sheffield, United Kingdom (virtual), 2021.

[8] N. Gutiérrez-Páez, J. S. Gómez-Cañón, L. Porcaro, P. Santos, D. Hernandez-Leo, and E. Gómez, "Emotion annotation of music: a citizen science approach," in *Proceedings of the 27th International Conference on Collaboration Technologies and Social Computing (CollabTech)*, Trier, Germany (virtual), 2021.

[9] Y.-H. Yang, Y.-F. Su, Y.-C. Lin, and H. H. Chen, "Music Emotion Recognition: The Role of Individuality," in *Proceedings of the International Workshop on Human-centered Multimedia (HCM)*, 2007, pp. 13–22.

[10] J. S. Gómez-Cañón, E. Cano, P. Herrera, and E. Gómez, "Joyful for you and tender for us: the influence of individual characteristics and language on emotion labeling and classification," in *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR)*, Montréal, Canada, 2020, pp. 853–860.

[11] A. Flexer and T. Grill, "The Problem of Limited Inter-rater Agreement in Modelling Music Similarity," *Journal of New Music Research*, vol. 45, no. 3, pp. 239–251, 2016.

[12] X. Hu and Y.-H. Yang, "Cross-dataset and cross-cultural music mood prediction: A case on Western and Chinese Pop songs," *IEEE Transactions on Affective Computing*, vol. 8, no. 2, pp. 228–240, 2017.

[13] D. Su and P. Fung, "Personalized music emotion classification via active learning," in *Proceedings of the 2nd International ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies*, New York, NY, USA, 2012, p. 57–62.

[14] B. Settles, *Active Learning*. Morgan and Claypool Publishers, 2012.

[15] Á. Sarasúa, C. Laurier, and P. Herrera, "Support Vector Machine Active Learning for Music Mood Tagging," in *Proceedings of the 9th International Symposium on Computer Music Modelling and Retrieval (CMMR)*, London, England, 2012, pp. 518–525.

[16] B. Settles, "Active learning literature survey," University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009.

[17] C. C. Aggarwal, X. Kong, Q. Gu, J. Han, and P. S. Yu, "Active Learning: A Survey," in *Data Classification: Algorithms and Applications*. CRC Press, 2014, pp. 571–605.

[18] Y. Yang, "Towards Practical Active Learning for Classification," Ph.D. dissertation, TU Delft University, 2018.

[19] P. Kumar and A. Gupta, "Active learning query strategies for classification, regression, and clustering: A survey," *Journal of Computer Science and Technology*, vol. 35, no. 4, pp. 913–945, Jul 2020.

[20] Y. Yan, R. Rosales, G. Fung, and J. G. Dy, "Active Learning from Crowds," in *Proceedings of the 28th International Conference on Machine Learning (ICML)*, Bellevue, USA, 2011.

[21] V. S. Sheng, F. Provost, and P. G. Ipeirotis, "Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers," in *Proceedings of the 14th International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD)*, Las Vegas, USA, 2008, pp. 614–622.

[22] P. Donmez, J. G. Carbonell, and J. Schneider, "Efficiently Learning the Accuracy of Labeling Sources for Selective Sampling General Terms," in *Proceedings of the 15th International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD)*, Paris, France, 2009, pp. 259–267.

[23] J. Zhang, X. Wu, and V. S. Sheng, "Learning from crowdsourced labeled data: a survey," *Artificial Intelligence Review*, vol. 46, pp. 543–576, 2016.

[24] Y.-A. Chen, J.-C. Wang, Y.-H. Yang, and H. Chen, "Linear Regression-based Adaptation of Music Emotion Recognition Models for Personalization," in *Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2014, pp. 2149–2153.

[25] Y.-A. Chen, J.-C. Wang, Y.-H. Yang, and H. H. Chen, "Component tying for mixture model adaptation in personalization of music emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1409–1420, 2017.

[26] D. Cohn, L. Atlas, R. Ladner, and A. Waibel, "Improving Generalization with Active Learning," *Machine Learning*, vol. 15, pp. 201–221, 1994.

[27] A. Flexer and T. Lallai, "Can we increase inter- and intra-rater agreement in modeling general music similarity?" in *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019, pp. 494–500.

[28] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Developing a benchmark for emotional analysis of music," *PLoS One*, pp. 1–22, 2017.

[29] J. A. Russell, "A circumplex model of affect," *Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[30] R. Panda, R. M. Rui, and P. Paiva, "Musical texture and expressivity features for music emotion recognition," in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2018.

[31] P. N. Juslin, *Handbook of Music and Emotion: Theory, Research, Applications*. Oxford: Oxford University Press, 2010.

[32] T. Eerola and J. K. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music," *Psychology of Music*, vol. 39, no. 1, pp. 18–49, 2011.

[33] T. Eerola, "Music and emotion," in *Handbook of Systematic Musicology*, R. Bader and S. Koelsch, Eds. Springer, 2018, ch. Music and Emotion, pp. 539–556.

[34] P. N. Juslin, *Musical Emotions Explained*. Oxford: Oxford University Press, 2019.

[35] Y. Chen, Y. Yang, J. Wang, and H. Chen, "The AMG1608 dataset for music emotion recognition," in *Proceedings of the 40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 693–697.

[36] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, K. R. Scherer, and J. Krajewski, "On the acoustics of emotion in audio: what speech, music, and sound have in common," *Frontiers in Psychology*, vol. 4, pp. 1–12, 2013.

[37] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent Developments in OpenSMILE, the Munich Open-source Multimedia Feature Extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, New York, NY, USA, 2013, pp. 835–838.

[38] M. Won, A. Ferraro, D. Bogdanov, and X. Serra, "Evaluation of cnn-based automatic music tagging models," in *Proceedings of 17th Sound and Music Computing Conference (SMC)*, 2020.

[39] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th Python in Science Conference (SCIPY)*, Austin, USA, 2015, pp. 18–25.

[40] H. Zhang, L. Jiang, and J. Su, "The optimality of naive bayes," in *Proceedings of the 17th Florida Artificial Intelligence Research Society Conference*, 2004, pp. 562–567.

[41] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, Aug 2016.

[42] L. Bottou, "Large-Scale Machine Learning with Stochastic Gradient Descent," in *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT)*, 2010, pp. 177–186.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[44] M. Won, S. Chun, and X. Serra, "Toward interpretable music tagging with self-attention," *arXiv preprint arXiv:1906.04972*, 2019.

[45] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[46] K. Bullard, Y. Schroecker, and S. Chernova, "Active Learning within Constrained Environments through Imitation of an Expert Questioner," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.

[47] S. Yang, M. Barthet, and E. Chew, "Multi-scale analysis of agreement levels in perceived emotion ratings during live performance," in *Late-Breaking Demo of the 18th International Society for Music Information Retrieval (ISMIR)*, Suzhou, China, 2017.

[48] K. H. Krippendorff, *Content Analysis: An Introduction to Its Methodology*, 2nd ed. SAGE Publications, 2004.

[49] M. Schedl, E. Gómez, E. S. Trent, M. Tkalcic, H. Eghbal-Zadeh, and A. Martorell, "On the interrelation between listener characteristics and the perception of emotions in Classical orchestra music," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 507–525, 2018.

[50] E. Bigand and J.-J. Aucouturier, "Seven problems that keep MIR from attracting the interest of cognition and neuroscience," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 483–497, 2013.

[51] H. V. Koops, W. Bas De Haas, J. A. Burgoyne, J. Bransen, A. Kent-Muller, and A. Volk, "Annotator subjectivity in harmony annotations of popular music," *Journal of New Music Research*, vol. 48, no. 3, pp. 232–252, 2019.

[52] A. Holzapfel, M. E. P. Davies, J. R. Zapata, J. L. Oliveira, and F. Gouyon, "Selective sampling for beat tracking evaluation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2539–2548, 2012.

[53] L. F. Barrett, *How Emotions are Made: The Secret Life of the Brain*. Houghton Mifflin Harcourt, 2017.