Harnessing Label Uncertainty to Improve Modeling: An Application to Student Engagement Recognition

Arkar Min Aung
Worcester Polytechnic Institute
Email: aaung@wpi.edu

Jacob R. Whitehill
Worcester Polytechnic Institute
Email: jrwhitehill@wpi.edu

***Abstract*—Automatic facial expression recognition systems are usually trained from target labels that model each example as belonging *unambiguously* to a single class (e.g., "non-engaged", "very engaged", etc.). However, in some settings, ground-truth labels can be more aptly modeled as probability distributions (e.g., $[0.1, 0.1, 0.5, 0.3]$ over 4 engagement categories) that capture the *uncertainty* that can arise during the annotation process. In this paper, we explore how harnessing the full probability distribution of each label ("soft labels"), rather than just a scalar summary statistic ("hard labels", e.g., majority class or mean), can yield better recognition accuracy when training automated detectors. Our results on a face image dataset (10698 faces over 20 subjects) labeled for perceived student engagement suggest that training on soft labels can deliver engagement detectors that fit the data stat. sig. more accurately (lower cross-entropy for classification, higher Pearson correlation for regression) than when training on hard labels. Moreover, we explore possible reasons for this effect and provide evidence that it is due to implicit regularization that the soft labels enact on the trained engagement detector. This effect is similar to, but empirically seems stronger than, the "label smoothing" approach proposed by Szegedy, et al. [1].**

Keywords*—*data annotation, label regularization, automatic facial expression recognition, student engagement recognition

I. Introduction

In automatic facial expression recognition and other affective computing applications, the predominant machine learning paradigm is to assign every example (e.g., an image or video of a person's face) a single, *unambiguous* ground-truth label representing the quantity that an automatic detector should predict. For an emotion recognition system, the label might be an element of a set of mutually exclusive basic emotions (anger, fear, joy, etc.); for a smile detector, it might simply be 1 or 0 to signify smile versus non-smile. These labels, along with the features constituting the examples themselves (e.g., image pixels), are then fed to an optimization algorithm (e.g., stochastic gradient descent) to train an automated detector.

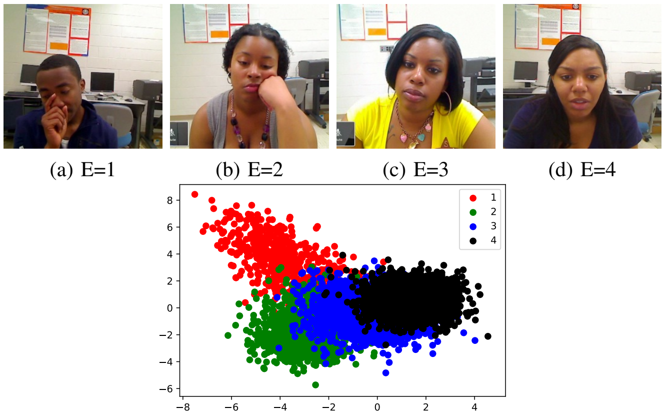
Multiple annotations per example: When collecting training and testing data, machine learning practitioners often collect multiple labels for each example, either from in-house annotators or using crowdsourcing, in order to improve labeling accuracy. The resulting distribution of labels for each example is then usually distilled into a single *summary statistic* that becomes the ground-truth for that image/video [2], [3], [4]. The summary statistic can be computed using a simple function such as majority vote or a more sophisticated consensus algorithm [5], [6], [7], [8]. Though the resulting label is aggregated from multiple, possibly differing opinions, it is treated as an *unambiguous* ground-truth label. For example, an image that is labeled by 4 labelers as "smile" and 1 labeler as "non-smile" might be assigned a ground-truth of 1 (smile). This ground-truth value could then be encoded in a 2-dimensional 1-hot vector (e.g., $[0, 1]$) and then used for training a neural network.

While simple and appropriate for certain problem domains, the "unambiguous label" paradigm misses an opportunity to model the *uncertainty* of each example's ground-truth label. Such uncertainty can arise due to *inherent* ambiguity in the perceptual process: for example, partial occlusion of the face might make it impossible for any external observer, no matter how skilled, to distinguish between two or more facial expression classes. It can also arise due to *subjective interpretation*: for example, annotators who are asked to rate how physically attractive they find a person portrayed in an image may disagree significantly. Some psychologists [9] argue that even the basic emotions [10] can be interpreted differently based on the context in which they are shown. The existence of label uncertainty raises the question: might training on the *full probability distribution* representing each label be better, in terms of downstream recognition accuracy, compared to the standard approach of training on only a summary statistic?

Student engagement recognition: As a motivating example, as well as the application focus of this paper, suppose we are training a detector that can analyze a frontal image of a student's face and estimate how *engaged* the student appears to be to an external observer (see Fig. 1 (top)). In this scenario (from [11]), there are four engagement categories, where 1 is least engaged and 4 is most engaged. Suppose that, when labeling a particular face image, one labeler assigns a label of E=1 and another labeler assigns a label of E=3. This could happen, for example, if the face image had one attribute (e.g., eye closure) that was associated with E=1, but also another attribute (e.g., in-plane rotation of the face) that was associated with E=3, and each labeler attended to only one of these two attributes. In this case, the probability distribution of ground-truth could be represented as $[0.5, 0, 0.5, 0]$. Is it reasonable in this case to summarize these two labels with their mean, i.e., E=2, and to train a classifier using this label (which would be 1-hot encoded as $[0, 1, 0, 0]$)? What if most faces in the E=2 category actually look quite different and possess *neither* of the properties in E=1 and E=3 images?

The hypothesis that summarizing a label distribution for a particular image could result in a ground-truth value that

This material is based upon work supported by the National Science Foundation under Grant No. #1551594.



(a) E=1 (b) E=2 (c) E=3 (d) E=4

Fig. 1: **Top**: Samples from the HBCU dataset [11], along with the levels of perceived *student engagement*, as rated by external coders. E=1 is the lowest engagement; E=4 is the highest. **Bottom**: Linear Discriminant Analysis of the HBCU faces. The average of the centroids of classes E=1 and E=3 in latent space does not correspond to the centroid of E=2.

fundamentally *mischaracterizes* the example and leads the classifier astray during training, is supported by a linear discriminant analysis we conducted on the faces in the HBCU student engagement dataset [11]: In Fig. 1 (bottom), we projected each face image into the two dimensions in which the four engagement levels can best be discriminated. It is evident that the average of the centroid of E=1 faces (red dots) and the centroid of the E=3 faces (blue dots) in the latent space does not correspond to the centroid of E=2 faces (green dots). This suggests that it might be better, when building an automatic engagement recognition system, to train on target values that capture the entire distribution of labels for each example, rather than summarizing it with a single value.

Contributions: In this paper we investigate the potential benefits, in terms of recognition accuracy, of training a facial expression recognition system using "soft" labels that capture the uncertainty of an image's ground-truth label, compared to the standard approach of using "hard" labels that distill each label distribution into a summary statistic. The context of our investigation is automatic student engagement recognition, which has garnered significant attention within the affective computing and intelligent tutoring systems communities in recent years [12], [11], [13], [14]. We present empirical results suggesting that soft labels deliver more accurate detectors, and we provide evidence that this accuracy improvement is due to a regularization effect implicitly induced by the soft labels.

II. Related Work

The issue of label uncertainty, how to model it, and how the label representation affects the recognition accuracy of trained classifiers has garnered a modest amount of research attention during the past 10 years. One of the two main approaches to dealing with multiple annotations per example is to employ a statistical *consensus algorithm* (e.g., [5], [6], [7], [8]) to aggregate multiple annotations into a single ground-truth label prior to classifier training. The other is to use a learning framework that can directly harness the entire *distribution* of labels collected from multiple labelers for each example; examples of such frameworks include probabilistic graphical models [15], fuzzy support vector machines [16], deep neural networks [17], and multi-score learning methods [18].

In terms of direct empirical comparison between soft and hard labels for classifier training, we are aware of only a few prior works: Scherer, et al. [16] found that fuzzy-input, fuzzy-output (F^2) support vector machines (SVMs) [19] trained on soft labels outperformed a standard RBF SVM trained on hard labels in a task on automatic speech analysis for voice attributes. Guan, et al. [17], in a study on automatic diagnosis of diabetic eye disease from retinopathy images, compared several neural network architectures designed to capture either the consensus label or the full distribution of labels; recognition accuracy of the best network trained to predict the full label distribution was slightly higher (AUC=0.9745) compared to a network to predict only the consensus label (AUC=0.9711). Nguyen, et al. [20] showed that using binary classification models enriched by auxiliary soft-label information outperforms traditional binary classifiers for predicting which patients are at risk for Heparin Induced Thrombocytopenia. To our knowledge, the only prior work that explicitly compared soft labels to hard labels for automatic face analysis was by Wang & Jinbo [21]: They trained a face classifier to discriminate joy from non-joy by formulating the optimization as a bi-convex program, based on a hinge loss between the model's predictions and the ground-truth labels. The classifier is binary and assumed to be linear, and it is not clear how their work could be extended to handle > 2 classes or non-linear classifiers. Moreover, their approach tries to assess the reliability of each labeler, and then to estimate the label of each image as a weighted combination of the labelers' opinions. In contrast, our paper explores the advantage of harnessing label uncertainty without inferring labelers' reliabilities.

III. Definitions

We define the **soft label** of an example (e.g., an image or video) to be the entire probability distribution of labels assigned to the example by a set of labelers. We define the **hard label** to be a scalar summary statistic – specifically, the rounded mean – of the soft label distribution.

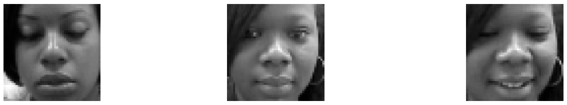
Example: Suppose some face image has been labeled by two labelers as engagement level 3 and by three labelers label as engagement level 4. Then the soft label of that face image would be represented by vector $l^s = [0, 0, 0.4, 0.6]$, where the ith component of l^s represents the probability (over all labelers for the image) that the image is labeled with engagement level i. The hard label for the image, on the other hand, would be the mean of this distribution (3.6) rounded to the nearest integer (4), and expressed as a 1-hot vector $l^h = [0, 0, 0, 1]$.

IV. Experiment I: Classification

We assessed how the type of the *training* labels (hard or soft) impacts the accuracy of an automatic student engagement *classifier* (over 4 engagement classes) trained with those labels.

A. Dataset

We used 10698 faces from the HBCU dataset [11] to run experiments. This dataset contains face images of 20 different



$l^s = [1,0,0,0]$
$l^h = [1,0,0,0]$

$l^s = [0,0,0.6,0.4]$
$l^h = [0,0,1,0]$

$l^s = [0.57,0,0.14,0.29]$
$l^h = [0,1,0,0]$

Fig. 2: Example faces and their soft (l^s) and hard (l^h) engagement labels, computed over the set of labelers who labeled each face.

African-American undergraduate students engaged in an educational game, along with labels of perceived "engagement" assigned to each face by multiple labelers. Each engagement label ranges from 1 to 4 (see Fig. 1 (top)). There were 7 unique labelers in total, such that: $< 1\%$ of images are labeled by 2 labelers; 56% by 3 labelers; 16% by 4 labelers; 24% by 5 labelers; 2% by 6 labelers and $< 1\%$ by 7 labelers. From each image, an automatic face detector [22] was used to crop the face and scale it to 48×48 face pixels (grayscale). See Fig. 2 for examples of the cropped faces and associated labels.

B. Architecture

We adopted the same Gabor + LogisticRegression architecture that was used for automatic student engagement recognition in [11]. The classifier is equivalent to a 3-layer neural network: The input layer consists of 48×48 grayscale pixel values. The first hidden layer is convolutional (40 feature maps with 48×48 kernels) and uses pre-computed, complex-valued weights (i.e., they are not trained) to compute Gabor Energy Filter [23] response values (5 spatial frequencies and 8 orientations spaced at $\pi/8$ radians). The non-linear activation function of this layer computes the absolute value of the complex-valued filter responses. The final layer is fully connected with 4 softmax outputs (one for each engagement level). Although the network's output is probabilistic, it can be converted into a "hard" label by taking the *rounded mean* of the predictive distribution. All network weights were initialized using the Xavier method [24] and optimized using the Adam optimizer with a learning rate of 0.001 for 45 training epochs.

Gabor-based face representations were, until the renaissance of deep neural networks, considered state-of-the-art for automatic facial expression recognition [23], [22], [25]. Interestingly, we found that, on the HBCU dataset, the Gabor-based network delivered *higher* performance and *lower* variance than the deeper architectures that we tried. Our hypothesis is that this is due to the small number of subjects in the HBCU dataset (only 20). See Supplementary Materials for more details.

C. Methods

We partitioned the faces from the HBCU dataset into 4 *subject-independent* cross-validation folds (the same used in [11]), where each fold contained 5 subjects. For each fold, we re-initialized the weights and trained the network in 25 different trials, and then averaged the results across these trials for each fold. We assessed accuracy using two metrics: cross-entropy loss, and percent-correct (equal to 1 minus the error rate). Moreover, we measure the accuracy of the classifier's predictions both w.r.t. both hard and soft *validation* labels.

Training Labels	Validation Labels	
	Hard	Soft
Hard	CE: 1.173 PC: 56.41%	CE: 1.587 PC: –
Soft	CE: 0.943 PC: 57.522%	CE: 1.103 PC: –

TABLE I: Results showing cross-entropy (**CE**) loss and percent-correct (**PC**) classification accuracy of facial engagement classifiers trained on either "hard" or "soft" labels, and evaluated on either "hard" or "soft" labels. **Trend**: Training on soft labels outperforms training on hard labels, both when validating on hard and on soft labels.

D. Results & Discussion

Results are shown in Table I. **Cross-entropy (CE) loss**: For all four folds (averaged over the 25 trials), training on soft labels resulted in lower (better) cross-entropy compared to training on hard labels; the difference was statistically significant and persisted both when validating on soft labels ($t(3) = 6.537$, $p = 0.007$, 2-tailed) and when validating on hard labels ($t(3) = 3.592$, $p = 0.037$, 2-tailed). Since cross-entropy is equivalent to the negative log-likelihood of the model predictions given the validation labels, this result indicates that training with soft labels results in a better model-fit (higher likelihood) on validation data. **Percent-correct (PC) classification accuracy**: PC was calculated by taking the rounded mean of each predictive distribution. (One could alternatively use the *argmax*, but empirically we found the results were worse, both for soft and hard labels.) A slight improvement in classification accuracy is observed when trained on soft labels, but the improvement was not statistically significantly different ($t(3) = -1.318$, $p = 0.279$, 2-tailed).

V. Experiment II: Regression

In the experiment above, a student's engagement was *classified* into the set $\{1,2,3,4\}$. Alternatively, it can be *regressed* into a real number (e.g., 3.2). In this section we measure the impact of soft versus hard labels on *regression* accuracy.

A. Architecture

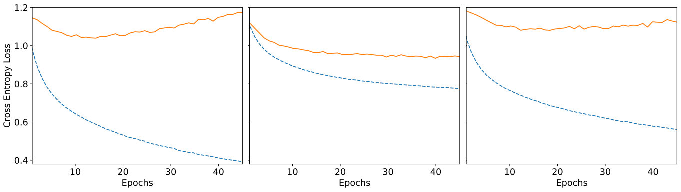
After training a 4-way engagement *classifier* as described above, we constructed a *regressor* by appending to the neural network a fourth layer with fixed (not optimized) weights that computes the inner product between the softmax engagement class probabilities $\{p_i\}$ and the vector $[1,2,3,4]$; the result ($\sum_{i=1}^{4} ip_i$) is the expected engagement level given the input image. We created regressors in this way for each of the two training label types (soft and hard). In addition, we also implemented the regression method used in [11], which directly estimates the engagement value from the Gabor layer without an intermediate softmax layer; we call this method "Gabor Direct". We trained this regressor by minimizing the squared-error loss w.r.t. the rounded mean engagement labels.

B. Methods

We used the same cross-validation methodology and folds as in Experiment I. To measure regression accuracy, we used Pearson's correlation of the regressor's prediction with respect to the rounded mean engagement level for each image.

	Training Approach			
	Hard	Soft	Gabor Direct [11]	Smoothed [1]
Pearson r	0.539	0.584	0.528	0.552

TABLE II: Pearson correlation coefficients, using different training approaches, of the predicted versus ground-truth (rounded mean) engagement labels.



(a) Hard (b) Soft (c) Smoothed [1]

Fig. 3: Training (blue dashed) & validation (orange solid) cross-entropy values over the 45 training epochs, when training on hard, soft, or smoothed (Sec. VI-A) labels. All validation losses are computed w.r.t. hard labels.

C. Results & Discussion

Results are shown in Table II. The Pearson correlation when training on soft labels ($r = 0.584$) is statistically significantly higher compared to training on hard labels ($r = 0.539$) ($t(3) = -3.661$, $p = 0.0352$, 2-tailed). It is also higher than the Gabor Direct method from [11], though the difference is not statistically significant ($t(3) = -1.359$, $p = 0.267$, 2-tailed). This suggests the better model-fit described in Sec. IV-D can also result in better regression accuracy.

VI. General Discussion

Why does training on soft labels result in higher accuracy? One possible reason is that modeling each example with the full distribution avoids *mischaracterizing* the data using a summary statistic (see Sec. I and Fig. 1 (bottom)). Another reason, which we investigate in this section, is that training on soft labels has a *regularization effect*: Optimizing the detector to emulate a *probability distribution* over all 4 engagement categories, rather than just a single unambiguous label (i.e., delta distribution), might encourage the classifier not to become too confident on training data and thus to avoid overfitting. Indeed, we find significant evidence to support this hypothesis in (1) the trajectories of cross-entropy values during training; and (2) the entropy of the detector's outputs on validation data.

Cross-entropy trajectories: Fig. 3 (a) shows the trajectory, over the 45 training epochs, of the average (over 4 folds and 25 trials/fold) cross-entropy loss values on training data (blue dashed line) and validation data (orange solid line), when training on *hard* labels. Fig. 3 (b) shows the analogous curves when training on *soft* labels. (All of the subplots in Fig. 3 show validation costs w.r.t. hard labels, but results are similar for the validation costs w.r.t. soft labels.) The plots show that, when training on hard labels, the *training* loss is lower than when training on soft labels, but the *validation* loss is higher. Moreover, the validation loss begins to increase around 20 epochs using hard labels, whereas it continues to decrease throughout the 45 training epochs using soft labels. Finally, the difference between training and validation curves is larger when training with hard labels than with soft labels.

Entropy of predictions: One of the symptoms of overfitting is that the classifier is too confident in its predictions, i.e., the average entropy of the output distributions is low. We found that the average entropy, on validation data, of the classifier that was trained on hard labels was 0.5578, whereas the entropy of the classifier trained on soft labels was higher: 0.8985. These results are all consistent with the hypothesis that soft labels provide a form of regularization.

A. Comparison to label smoothing

Szegedy, et al. [1] introduced a form of regularization that operates by "smoothing" the input labels: in particular, the original training label distribution for each example is replaced with a mixture model consisting of the original ground-truth (delta distribution) label (with weight $1 - \epsilon$) and the uniform distribution (weight ϵ). Might this have a similar effect as soft labels? To explore this question, we trained an engagement *classifier* using smoothed labels where the ϵ was optimized using a grid search (over the interval $[0.02, 0.2]$). To give the smoothed labels approach the best chance of succeeding, we chose ϵ to minimize the cross-entropy directly on the *validation set* (with hard test labels). Even for the best ϵ value (which was 0.06), the cross-entropy at the end of 45 training epochs was higher (worse) using smoothed labels than with soft labels (see Fig. 3 (c)). We do observe that smoothed labels have a regularization effect, i.e., smaller gap between training and validation cross-entropy losses compared to training on hard labels; however, the effect is less pronounced than when training on soft labels. The average entropy on validation data of the estimated class probabilities $\{p_i\}$ for each example is higher using smoothed labels (0.7259) than with hard labels (0.5578), but smaller than with soft labels (0.8985). Altogether, these results suggest that soft and smoothed labels have a similar effect, but soft labels have a more pronounced effect. A possible explanation for this difference is that soft labels capture the *actual* label distribution rather than adding an *artificial* distribution as in label smoothing.

As a final comparison between smoothed and soft labels, we created a *regressor* (as described in Sec. V-A) by appending an inner-product layer with fixed weights $[1, 2, 3, 4]$ to the classifier trained using smoothed labels. The Pearson correlation of the predicted engagement scores w.r.t. rounded mean ground-truth labels was higher (0.552) than using hard labels (0.539), but lower than with soft labels (0.584).

VII. Conclusion

We have provided evidence that training on "soft" labels can deliver more accurate automatic student engagement detectors, both in terms of classification and regression accuracy. Moreover, we have shown evidence that this improvement is due to regularization induced by soft labels. **Future work**: Our focus was on modeling uncertainty induced by differing opinions across multiple labelers. However, as mentioned in the introduction, uncertainty can also arise when examples are *inherently* ambiguous. In such cases, *each* label might consist of a probability distribution that expresses an *individual* labeler's uncertainty. It would be interesting to explore whether the advantage of soft labels persists in this setting.

References

[1] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

[2] F. Eyben, M. Wöllmer, M. F. Valstar, H. Gunes, B. Schuller, and M. Pantic, "String-based audiovisual fusion of behavioural events for the assessment of dimensional affect," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pp. 322–329, IEEE, 2011.

[3] G. Tavares, A. Mourão, and J. Magalhães, "Crowdsourcing facial expressions for affective-interaction," *Computer Vision and Image Understanding*, vol. 147, pp. 102–113, 2016.

[4] G. Chittaranjan, O. Aran, and D. Gatica-Perez, "Exploiting observers' judgements for nonverbal group interaction analysis," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pp. 734–739, IEEE, 2011.

[5] P. Welinder, S. Branson, P. Perona, and S. J. Belongie, "The multidimensional wisdom of crowds," in *Advances in neural information processing systems*, pp. 2424–2432, 2010.

[6] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *Journal of Machine Learning Research*, vol. 11, no. Apr, pp. 1297–1322, 2010.

[7] 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 *Advances in neural information processing systems*, pp. 2035–2043, 2009.

[8] D. Zhou, Q. Liu, J. Platt, and C. Meek, "Aggregating ordinal labels from crowds by minimax conditional entropy," in *International Conference on Machine Learning*, pp. 262–270, 2014.

[9] J. M. Carroll and J. A. Russell, "Do facial expressions signal specific emotions? judging emotion from the face in context.," *Journal of personality and social psychology*, vol. 70, no. 2, p. 205, 1996.

[10] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[11] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagementfrom facial expressions," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 86–98, 2014.

[12] N. Bosch, S. D'Mello, R. Baker, J. Ocumpaugh, V. Shute, M. Ventura, L. Wang, and W. Zhao, "Automatic detection of learning-centered affective states in the wild," in *Proceedings of the 20th international conference on intelligent user interfaces*, pp. 379–388, ACM, 2015.

[13] H. Monkaresi, N. Bosch, R. A. Calvo, and S. K. D'Mello, "Automated detection of engagement using video-based estimation of facial expressions and heart rate," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 15–28, 2017.

[14] N. Bosch, S. K. D'mello, J. Ocumpaugh, R. S. Baker, and V. Shute, "Using video to automatically detect learner affect in computer-enabled classrooms," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 6, no. 2, p. 17, 2016.

[15] S. Kumano, K. Otsuka, D. Mikami, M. Matsuda, and J. Yamato, "Analyzing interpersonal empathy via collective impressions," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 324–336, 2015.

[16] S. Scherer, J. Kane, C. Gobl, and F. Schwenker, "Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification," *Computer Speech & Language*, vol. 27, no. 1, pp. 263–287, 2013.

[17] M. Y. Guan, V. Gulshan, A. M. Dai, and G. E. Hinton, "Who said what: Modeling individual labelers improves classification," *arXiv preprint arXiv:1703.08774*, 2017.

[18] H. Meng, A. Kleinsmith, and N. Bianchi-Berthouze, "Multi-score learning for affect recognition: the case of body postures," in *International Conference on Affective Computing and Intelligent Interaction*, pp. 225–234, Springer, 2011.

[19] C. Thiel, S. Scherer, and F. Schwenker, "Fuzzy-input fuzzy-output one-against-all support vector machines," in *Knowledge-based intelligent information and engineering systems*, pp. 156–165, Springer, 2007.

[20] Q. Nguyen, H. Valizadegan, and M. Hauskrecht, "Learning classification models with soft-label information," *Journal of the American Medical Informatics Association*, vol. 21, no. 3, pp. 501–508, 2014.

[21] X. Wang and J. Bi, "Bi-convex optimization to learn classifiers from multiple biomedical annotations," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 14, no. 3, pp. 564–575, 2017.

[22] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, "The computer expression recognition toolbox (cert)," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pp. 298–305, IEEE, 2011.

[23] T. Wu, M. S. Bartlett, and J. R. Movellan, "Facial expression recognition using gabor motion energy filters," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pp. 42–47, IEEE, 2010.

[24] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256, 2010.

[25] A. B. Ashraf, S. Lucey, and T. Chen, "Reinterpreting the application of gabor filters as a manipulation of the margin in linear support vector machines," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 7, pp. 1335–1341, 2010.