Addressing Bias in Machine Learning Algorithms: A Pilot Study on Emotion Recognition for Intelligent Systems

Ayanna Howard[1*], Cha Zhang[2], Eric Horvitz[2]
[1]School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA

[2]Microsoft Research
Redmond, WA

Abstract: **Recently, there has been an explosion of cloud-based services that enable developers to include a spectrum of recognition services, such as emotion recognition, in their applications. The recognition of emotions is a challenging problem, and research has been done on building classifiers to recognize emotion in the open world. Often, learned emotion models are trained on data sets that may not sufficiently represent a target population of interest. For example, many of these on-line services have focused on training and testing using a majority representation of adults and thus are tuned to the dynamics of mature faces. For applications designed to serve an older or younger age demographic, using the outputs from these pre-defined models may result in lower performance rates than when using a specialized classifier. Similar challenges with biases in performance arise in other situations where datasets in these large-scale on-line services have a non-representative ratio of the desired class of interest. We consider the challenge of providing application developers with the power to utilize pre-constructed cloud-based services in their applications while still ensuring satisfactory performance for their unique workload of cases. We focus on biases in emotion recognition as a representative scenario to evaluate an approach to improving recognition rates when an on-line pre-trained classifier is used for recognition of a class that may have a minority representation in the training set. We discuss a hierarchical classification approach to address this challenge and show that the average recognition rate associated with the most difficult emotion for the minority class increases by 41.5% and the overall recognition rate for all classes increases by 17.3% when using this approach.**

I. INTRODUCTION

Poor representation of people of different ages and skin colors in training data can lead to performance problems and biases for real-world classification tasks involving the visual attributes of people—such as detecting facial expressions or pose. These performance problems and biases are directly correlated with the problem of class-imbalance in the datasets used to train these machine learning algorithms. There have been a number of efforts that have tried to resolve this issue with training on imbalanced datasets [1], including using different forms of re-sampling [2], adjusting the decision threshold [3], or a mixture-of-experts approach which combines the results of many classifiers [4]. The difficulty with imbalance is recognizing why and when imbalanced data sets are problematic. It can be difficult to distinguish between data associated with a low incidence class and noisy data when training a classifier. This is especially problematic when building models using cloud-based services that utilize training data drawn from readily available sources, such as photos crawled from the web. Tens-of-thousands of cases might be downloaded for such training, and the resulting datasets may be dominated by high prevalence age and skin color categories. Such broad scans of data lead to three challenges for application developers, including machine learning practitioners and roboticists. The first challenge is that, by using a cloud-based service, application developers typically cannot directly influence its behavior since they do not have access to the services' internal processes. The second challenge is that the categories which are derived from a representative quantity of majority data (and thus have likelihoods representative of real-world data streams) may lead to incorrect outcomes when a person in a minority age and skin color category uses the system. In this instance, the learning converges based on theoretically acceptable outcomes but its real-world outcomes may not be socially acceptable. The third challenge deals with the amplification of the problem when the data source is or is almost completely influenced by the dominant class, which may not be representative of the world culture, thus leading to the perpetuation of biased beliefs. Although the derivations of these problems are different, these situations can bias the classification results, especially when designing learning algorithms for emotion recognition in intelligent systems.

Emotion recognition is a growing area of interest, from interpreting moods for effective caregiver robots to recognizing a child's anxiety level for therapeutic robots. Emotion recognition is the process of identifying human emotion, typically via facial expressions, linguistics, voice, or even body gestures. Most machine learning algorithms for emotion recognition use images to track facial expressions in

[*] Research performed while Visiting Researcher at Microsoft Research

order to identify basic emotions such as Happy, Angry, Fear, or Surprise. In fact, in a review of online emotion recognition APIs [5], over 50% of the intelligent software packages available used facial expressions to recognize emotions.

Generalization of these algorithms for optimizing performance in the open world is typically achieved by training on large sets of unconstrained data collected 'in the wild'. The term 'in the wild' means images have been captured under natural conditions with varying parameters such as environments and scenes, diverse illumination conditions, head poses and with occlusions. Unfortunately, most of these image sets, by the nature of their collection process, will have small collections of low incidence classes, such as those associated with a younger or older age demographic, or with an ethnic minority. For example, in the yearly Emotion Recognition in the Wild (EmotiW) challenge [6], researchers are provided a dataset to benchmark the performance of their methods on in-the-wild data. The dataset represents the typical expression classes of Angry, Disgust, Fear, Happiness, Neutral, Sadness and Surprise, but focuses primarily on scenes with adults.

In this paper, we examine the bias issue found in learning algorithms for intelligent systems by focusing on the emotion recognition problem. We first present baseline outcomes for a cloud-based emotion recognition algorithm applied to images associated with a minority class, in this instance, children's facial expressions. We then present a hierarchical approach that combines outputs from the cloud-based emotion recognition algorithm with a specialized learner, and show that this methodology can increase overall recognition results by 17.3%. We also verify that this hierarchical algorithm, when applied to an additional corpus of test data, shows similar improvements in the recognition rate. This additional corpus is used to assess the performance of the hierarchical approach in generalizing to new unseen images. We conclude by discussing future work needed to address the problem of bias stemming from poor representation of people in a large training set.

II. Related Work

Although there are a number of research efforts focused on children that incorporate the recognition of emotions to enable their functionality, most have not done a systematic analysis of accuracy with respect to their emotion recognition algorithms. For example, with socially-interactive robots, a number of research robots such as Darwin [7], the iCub [8], Avatar [9], and the GRACE robot [10], use emotions to engage children in therapy or learning. Their analysis though is based on overall performance on child engagement and not on emotion recognition.

Of those research efforts that have focused on emotion recognition performance [11], very few have focused on children. In [12], researchers used their own video data collection of children to develop methods to analyze changes in facial expressions of 50 children age 3 to 9, with a focus on problem solving scenarios. Their accuracy measures were not directly based on emotions but rather centered on Facial Action Units, which, when blended together, can be used to represent emotions. In their application, they developed separate linear support vector machines to detect the presence or absence of one of 19 facial actions. The training set consisted of 10000 image frames and the testing set consisted of 200 randomly sampled frames from this set, resulting in a recognition rate for the facial action units that ranged from 61% to 100% depending on the unit. In [13], an approach for learning children's affective state was presented using three different types of neural network structures, namely a multi-stage radial basis function neural network, a probabilistic neural network, and a multi-class classification support vector machine (SVM). Using the Dartmouth Database of Children Faces [14], they subdivided images of children age 5 to 9 into a training set of 1040 images and a testing set of 242 images. The training set was clustered into three affective classes: positive (Happy, Pleased, Surprised), negative (Disgust, Sad, Angry) and Neutral. They achieved a maximum overall recognition rate of 85% on the untrained facial test images using the multi-class classification SVM. In [15], a method was discussed for automatic recognition of facial expressions for children. They validated their methodology on the full Dartmouth Database of Children Faces of 1280 images and their 8 emotions classes. The full image set was used for both training and testing using a support vector machine, a C4.5 decision tree, random forest and the multi-layer perceptron method. The SVM achieved the maximum overall recognition rate of 79%. They then tested on the NIMH child emotional faces picture set (NIMH-ChEFS) database [16] with 482 images to assess the generalization accuracy of the classifiers on new unseen images and achieved a maximum overall recognition rate of 68.4% when using the SVM that was trained on the Dartmouth dataset.

These represent several efforts that have focused on the explicit evaluation of emotion recognition algorithms as applied to the domain of children's emotional cues. In the next section, we discuss baseline results derived from using a cloud-based classifier on publically available datasets of children's facial expressions.

III. Baseline Results

Recently, several cloud-based services have offered programming libraries that enable developers to include emotion recognition capabilities in their imaging applications [17]. These include, among others, Google's Vision API (https://cloud.google.com/vision) and the Microsoft Emotion API (https://www.microsoft.com/cognitive-services), a component of Microsoft's Cognitive Services. These cloud-based emotion recognition algorithms optimize their performance in the open world by training on large sets of unconstrained data sets collected 'in the wild.' To establish a baseline on the capabilities of learning and inference for a minority class, we evaluate the emotion recognition results associated with children's facial expressions using the Microsoft Emotion API, a deep learning neural network [18]. The emotions detected using the Microsoft Emotion API are Angry, Contempt, Disgust, Fear, Happy, Neutral, Sad, and Surprise.

We selected four datasets of children's faces that are publically available for research purposes and with published

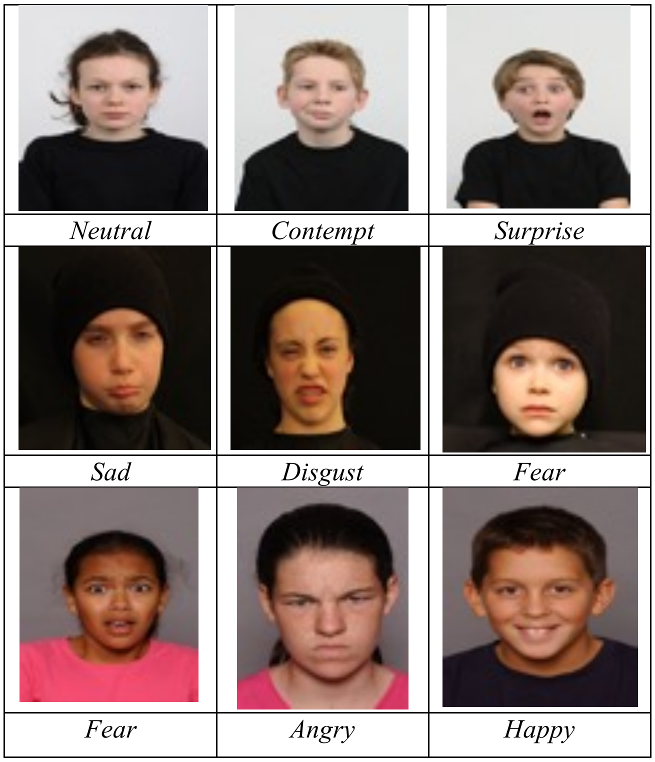


Fig. 1. Example stimuli of children associated with the facial expression databases: Top: The Radboud Faces Database, Middle: The Dartmouth Database of Children's Faces, Bottom: NIMH-ChEFS Database

human inter-rater reliability measures associated with the emotion labels. The four datasets were the NIMH Child Emotional Faces Picture Set (NIMH-ChEFS) [16], the Dartmouth Database of Children's Faces [14], the Radboud Faces Database [19], and the Child Emotions Picture Set (CEPS) [20] (Figure 1). The NIMH Child Emotional Faces Picture Set (NIMH-ChEFS) contains 534 images of children ranging in age from 10 to 17 years old with a mean age of 13.6 years old. The picture set includes 39 girls and 20 boys covering 5 emotions (Fear, Angry, Happy, Sad and Neutral). We selected an inter-rater reliability value for inclusion in our evaluation at 75% of the raters correctly identifying the intended emotion, which excluded 52 pictures from the original set leaving a final set of 482 pictures. Using inter-rater reliability for inclusion provides a way of quantifying the degree of agreement between two or more coders. We selected 75% as this is the benchmark established for having good agreement among raters when rating among 5-7 categories. The Dartmouth Database of Children's Faces contains 1280 images of children ranging in age from 5 to 16 years old with a mean age of 9.72 years old. The picture set includes 40 girls and 40 boys covering 7 emotions (Neutral, Happy, Sad, Angry, Fear, Surprise, and Disgust). For evaluation, we selected the inter-rater reliability cut-off value for inclusion at 75%, which excluded 370 pictures from the original set leaving a final set of 910 pictures. The Radboud Faces Database (RaFD) includes 240 images covering 8 emotions (Neutral, Angry, Sad, Fear, Disgust, Surprise, Happy, and Contempt). There are 4 boys and 6 girls in the dataset. Based on a 75% inclusion criteria, we excluded 57 pictures from the original set leaving a final set of 183 pictures. Lastly is the Child Emotions Picture Set (CEPS), which contains 273 images of children, ranging in age from 6 to 11 years old with a mean age of 8.9 years old. The dataset includes 9 girls and 8 boys covering 7 emotions (Happy, Sad, Angry, Disgust, Fear, Surprise, and Neutral). Since we did not have access to the individual inter-rater reliability values associated with this dataset, we used the researcher's inclusion critera, which resulted in a final set of 225 images.

We seek to characterize the performance of existing machine learning algorithms on cases from these distinct data sets to develop an understanding of how well the deep learning neural network (i.e. the Microsoft Emotion API) performs on recognizing children's emotional states.

To evaluate the performance of the deep learning algorithm, we parsed each of the final image sets into the Emotion API and tabulated the recognition results. Table I shows the performance for each of the datasets.

TABLE I. DEEP LEARNING RECOGNITION RATES ACROSS THE DIFFERENT STIMULI SETS (IN %): (FE)AR, (AN)GRY, (HA)PPY, (SA)D, (NE)UTRAL, (SU)RPRISED, (DI)SGUST, (CO)NTEMPT

	Fe	An	Di	Ha	Ne	Sa	Su	Co
NIMH-ChEFS	13	43		100	100	48		
Dartmouth	25	35	55	100	99	64	91	
Radboud	33	54	100	100	100	95	100	50
CEPS	5	50	10	95	92	52	81	

With respect to the overall recognition rates, which incorporate the variable data dimensions of the image sets (Table II), the overall emotion recognition rate was 62% for the NIMH-ChEFS dataset, 81% for the Dartmouth dataset, 83% for the Radboud dataset, and 61% for the CEPS dataset. If we compare these results with specialized learning algorithms that have been trained specifically on children's facial images and as discussed in the related work section, we see that the deep learning algorithm, based on training in the wild, has results that are comparable to the specialized results that used comparatively smaller training sets but with an emphasis on a minority class (Table III). Yet, if we look at the overall emotion recognition rates for adults, the rates should be closer to 88% [30]. Our goal therefore is to capitalize on the power of the cloud-based emotion recognition algorithm while improving the overall recognition rate.

TABLE II. NUMBER OF IMAGES ASSOCIATED WITH EACH STIMULI SET: (FE)AR, (AN)GRY, (HA)PPY, (SA)D, (NE)UTRAL, (SU)RPRISED, (DI)SGUST, (CO)NTEMPT

	Fe	An	Di	Ha	Ne	Sa	Su	Co
NIMH-ChEFS	102	94		108	98	80		
Dartmouth	20	83	101	302	145	135	124	
Radboud	21	26	25	30	24	20	29	8
CEPS	20	30	31	56	24	33	31	

TABLE III. DEEP LEARNING 'IN THE WILD' APPROACH VERSUS SPECIALIZED LEARNING METHODOLOGIES

	Dartmouth Database *with emotions grouped into positive, negative, and neutral affective states*	
Albu [13]	85%	
Emotion API	84%	
	Dartmouth Database	NIMH-ChEFS
Khan [15]	79%	68%
Emotion API	81%	62%

III. METHODOLOGY AND RESULTS

When we examine the results from the deep learning neural network, we note that Fear has a significantly lower recognition rate than the other emotions across all of the different datasets. If we look at the confusion matrix associated with Fear (Table IV), we also note that Fear is most often confused with Surprise. Facial expressions of Fear and Disgust have repeatedly been found in the emotion recognition literature to be less well recognized than those of other basic emotions [21]. In fact, it has been shown that Surprise is the most frequent error made when trying to recognize expressions of Fear in children [22]. If we look at the basic facial action units that make up these two expressions, it becomes obvious why this confusion occurs. Facial Action Units (AUs) represent the 44 anatomically distinct muscular activities that activate when changes occur in an individual's facial appearance (Table V). Based on comprehensive comparative human studies, Ekman and Friesen [23, 24] have labeled these muscle movements and identified those believed to be associated with emotional expressions (Table VI) [25]. When examining the facial action units associated with Fear and Surprise (Table V), we confirm that Surprise is actually a subset of Fear (i.e. Surprise ⊆ Fear); and over 60% of the AUs in Fear are also found in the set of Surprise AUs. As such, Surprise becomes easier to recognize than Fear as a default since there are less distinctive cues to identify.

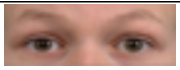
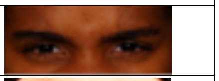
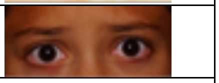
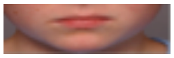
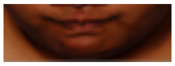
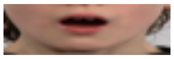
TABLE IV. CONFUSION MATRIX ASSOCIATED WITH DEEP LEARNING RESULTS

	Surprise	**Neutral**	**Happy**	**Sad**	**Fear**
Fear					
NIMH-ChEFS	74 (83.7%)	11 (10.2%)	0	3	13 (6.1%)
Dartmouth	11 (55%)	0	3 (15%)	1 (5%)	5 (25%)
Radboud	13 (61.9%)	0	0	1 (4.8%)	7 (33.3%)
CEPS	9 (45%)	4 (20%)	3 (15%)	3 (15%)	1 (5%)
Surprise					
Dartmouth	113 (91.1%)	3 (2.4%)	8 (6.5%)	0	0
Radboud	29 (100%)	0	0	0	0
CEPS	25 (80.6%)	3 (9.7%)	3 (9.7%)	0	0

TABLE V. FACIAL ACTION UNITS INVOLVED IN EMOTION STIMULI

Emotion	**AUs associated with Emotion**
Angry	4, 5 and/or 7, 22, 23, 24
Fear	1, 2, 4, 5, 7, 20, 25 or 26
Surprise	1, 2, 5, 25 or 26

TABLE VI. FACIAL ACTION UNITS AND FACIAL FEATURE IMAGES FROM THE CHILDREN FACIAL EXPRESSION DATASETS

Action Unit	**Description**	**Facial Feature Image**
1	Inner Brow Raiser	
2	Outer Brow Raiser	
4	Brow Lowerer	
5	Upper Lid Raiser	
7	Lid Tightener	
20	Lip Stretcher	
22	Lip Funneler	
23	Lip Tightener	
24	Lip Pressor	
25	Lips Apart	
26	Jaw Drop	

Thus, as a first step in illustrating a process for improving the recognition rates of a generalized machine learning algorithm, we focus on improving the overall recognition rates by improving recognition of the Fear emotion. Figure 2 depicts the overall algorithmic flow of the approach. Given a facial image, facial landmarks are extracted and used to compute a number of anthropometric features. The anthropometric features are then fed into two Support Vector Machines (SVMs) for binary classification, one to distinguish between Fear and Surprise with an explicit bias toward Fear and one to distinguish between Surprise and Not-Surprise with a balanced bias. The design of this construct is to increase the bias toward the minority class (in the first SVM) while ensuring that the recognition of the majority class is not drastically reduced (in the second SVM). We train the SVMs on 50% of the data from three of the datasets of children's faces and evaluate the results on all four datasets, including the remaining untrained dataset.

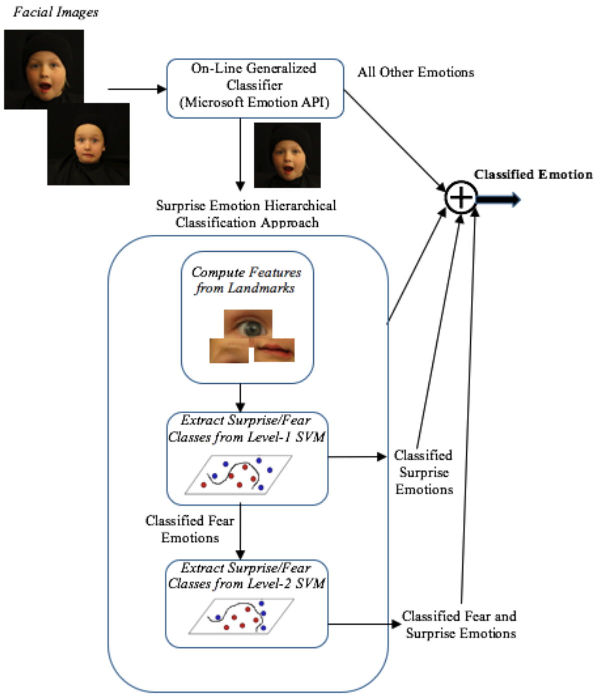

Facial Images
On-Line Generalized Classifier (Microsoft Emotion API)
All Other Emotions
Classified Emotion
Surprise Emotion Hierarchical Classification Approach
Compute Features from Landmarks
Extract Surprise/Fear Classes from Level-1 SVM
Classified Surprise Emotions
Classified Fear Emotions
Extract Surprise/Fear Classes from Level-2 SVM
Classified Fear and Surprise Emotions

Fig. 2. Feature-based learning approach for emotion recognition of children's facial expressions

A. Extraction of Anthropometric Features

Most methods that focus on developing image-based age classification methods typically use features associated with face anthropometry [26]. A face anthropometric model is based on measurements of size and proportions of the human face, i.e. human face ratios. Although age classification is not our direct target application, it does provide some measure for distinguishing between age groups (i.e. children versus adults) based on facial features. We thus utilize various human face ratios as the input into our specialized learning algorithm. In [27], it was shown that four feature distances are sufficient to represent the mathematical relationships among all of the various landmarks for age classification. Since we are interested in emotion classification, we compute all principal ratios, namely: Width of Left Eye, Height of Left Eye, Length of the Nose Bridge, Distance between the Nose Tip and Chin, Width of Mouth, Height of Open Mouth, and Offset Distance between the Inner and Outer Corner of the Eyebrow. These anthropometric features are computed based on extracted face landmarks, as shown in Figure 3, and Equations (1)-(7).

$$WidthofLeftEye = |EyeLeftInner_x - EyeLeftOuter_x| \quad (1)$$

$$NoseBridgeLength = |NoseRootRight_x - NoseRootLeft_x| \quad (2)$$

$$NoseTiptoChin = |ChinPosition_{\,y} - NoseTip_y| \quad (3)$$

$$HeightofLeftEye = |EyeLeftBottom_y - EyeLeftTop_y| \quad (4)$$

$$WidthofMouth = |MouthRight_x - MouthLeft_x| \quad (5)$$

$$HeightofOpenMouth = |UpperLipBottom_y - UnderLipTop_y| \quad (6)$$

$$EyeBrowHeightOffset = |EyebrowLeftInner_y - EyebrowLeftOuter_y| \quad (7)$$

Where x and y represent the pixel location in (x,y) screen coordinates and ChinPosition is estimated as the pixel coordinates associated with the bottom center of the Face Rectangle provided by the Face API, which indicates where in the image a face is located. Once computed, all ratios are normalized based on the calculated width and height of the Face Rectangle.

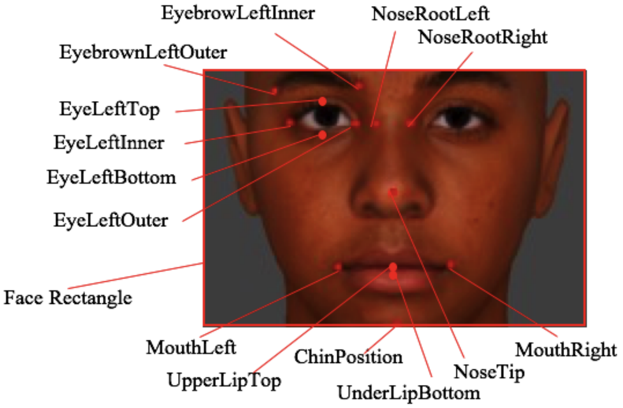


Fig. 3. Face Landmarks extracted using the Face API developed by Microsoft Oxford Project[1]

Once computed, these features are used to train two SVMs for emotion classification.

B. SVMs for Classification of Fear versus Surprised

Given that there is a bias for Surprise versus Fear associated with children's facial expressions, our first task was to develop a specialized classifier that biases the results toward the minority class, in this case, the Fear class. Support Vector Machines (SVMs) are supervised learning models that can be used for classification of classes associated with a set of training examples [28]. For our application, we want to better differentiate between the Fear minority class and the Surprise majority class. Thus, any facial expression that was classified (correctly or incorrectly) as Surprised by the deep learning algorithm, we want to re-evaluate with a specialized learner. To enable this process, all emotions labeled as Surprise are fed to the first-level SVM and reclassified into one of two classes: Fear or Surprise. We thus design the first-level SVM to learn the mapping:

$$X \mapsto Y, \text{where } x \in R^n, y \in \{\pm 1\}, n = 7 \quad (8)$$

In this case, *x* represents the vector containing anthropometric features, *y=1* represents the Surprise class, and *y=-1* represents the Fear class.

[1] https://www.projectoxford.ai/face

For our application, we trained the first-level SVM on 50% of the feature vectors classified as Fear or Surprise by the deep learning algorithm and extracted from the Radboud Faces Database, the Dartmouth Database of Children's Faces and the NIMH-ChEFS Database. We did not train on the Child Emotions Picture Set as we wished to assess the capabilities of the new algorithm when faced with unseen facial characteristics. We then biased the class decision threshold of the first-level SVM by selecting the minimum threshold value that maximized the true positive rate associated with Fear. This, by default, increases the false positive rate of Fear while potentially reducing the true positive rate associated with Surprise. Thus, after parsing the images through the first-level SVM, the minority class has a significantly higher recognition rate but the majority class recognition rate, on average, is reduced as shown in Table VII. The higher recognition rates for the minority class are valid even for the CEPS database which contains data that was not in the training sets of of the SVM classifiers. All recognition rates incorporate the variable data dimensions of the image sets (Table II).

TABLE VII. EMOTION RECOGNITION RATES AFTER TRAINING: ML – DEEP LEARNING ALGORITHM, SVM – FIRST-LEVEL SUPPORT VECTOR MACHINE

	Fear		**Surprise**		**Change in Overall Rec. Rate**
	ML	**ML+ SVM**	**ML**	**ML+ SVM**	
NIMH-ChEFS	13%	47%			34%
Dartmouth	25%	91%	91%	79%	-1.2%
Radboud	33%	77%	100%	100%	31.8%
CEPS	5%	70%	81%	68%	17.6%

The goal of the second-level SVM is to increase the recognition rate of the majority class to pre-bias levels while still keeping the recognition rate associated with the minority class higher than its original recognition rate. From Table V, we note that Angry and Fear have more Action Units in common than Angry and Surprise. Thus, to reduce the effect of the Action Unit overlap between Surprise and Fear, we trained the second-level SVM on the recognition of two primary classes – (Fear ∨ Angry) and Surprise. We then associated the derived anthropometric feature vector from each image to one of two classes: Surprised and Not-Surprised, where Not-Surprised represented the union of Fear and Angry. In this case, for the mapping: $X \mapsto Y$, where $x \in R^n, y \in \{\pm 1\}, n = 7$, *y=1* is associated with the Surprise class, and *y=-1* is associated with the Fear or Angry class. In practice, only those feature vectors that were classified as Fear by the first-level SVM are processed by the second-level SVM. This approach results in an average increase in the recognition of Fear by 41.5% and an increase in the overall recognition rate by 17.3%. As Table VIII shows, the overall recognition rate increases after parsing through the second-level SVM, even though the recognition rate for the minority class falls to a lower value than in the first-level SVM. Of special interest, we note that, although recognition of Fear has increased greatly for the Dartmouth database, the recognition rate for Surprise is slightly lower than the original Surprise recognition rate, even after parsing through the second-level SVM. Of all the datasets, the Dartmouth dataset had a large imbalance between Surprise and Fear, in fact this dataset had 6x more images belonging to Surprise than Fear (Table II). As such, this result is not surprising as the balance that we are trying to achieve in our approach is to ensure that the minority class has an increase in benefit with respect to recognition rates, while ensuring that the reduction in benefits to the majority class is not major. If the motivation is, instead, to ensure that the recognition rate for the minority class is maximized, the only requirement is to ignore the second-level SVM and utilize only the output resulting from parsing through the first-level SVM.

In the next section, we make some interesting observations about these results and provide discussion on ways to generalize this approach to the broad class of generalized learning algorithms.

TABLE VIII. EMOTION RECOGNITION RATES AFTER TRAINING: ML – DEEP LEARNING ALGORITHM, SVM – SECOND-LEVEL SUPPORT VECTOR MACHINE

	Fear		**Surprise**		**Change in Overall Rec. Rate**
	ML	**ML+ SVM**	**ML**	**ML+ SVM**	
NIMH-ChEFS	13%	47%			34%
Dartmouth	25%	70%	91%	83%	-0.6%
Radboud	33%	71%	100%	100%	32.8%
CEPS	5%	55%	81%	81%	19.6%

IV. DISCUSSION

We conclude this paper with a discussion on the presented results and highlight some areas for future efforts that could address the limitations associated with building classifiers when there is imbalanced representation in their training sets.

Recently, there has been an upsurge of attention given to generalized machine learning algorithms and the practices of inequality and discrimination that are potentially being built into them [29]. We know that imbalances exist and thus, our goal in this paper is to present an approach that enables us to capitalize on the power of generalized learning algorithms, while incorporating a process that allows us to tune those results for different target demographics. Bias in machine learning algorithms will occur anytime there is a large majority class coupled with other minority classes having lower incidence rates, such as those associated with a younger or older age demographic, or an ethnic minority. The challenge is to develop a process for ensuring the overall positive results of the generalized learning approach is maintained, while also increasing the outcomes associated with any minority classes. In this paper, we address this issue by developing a hierarchical approach that couples the results from the generalized learning algorithm with results from a specialized learner. Although we focus on the issue of emotion recognition for intelligent systems, and address emotion recognition associated with children's facial expressions, this concept can be applied to similar classification applications. The steps involved are (1) identifying the set(s) of minority classes, (2) developing specialized learners that address the minority class via special focus on the class, and (3) developing a specialized learner that combines signals from both the minority and majority class models.

As shown in the results, if at any point, we determine that it is more important to have a maximum outcome rate associated with the minority class, regardless of the outcome rate associated with the majority class, only steps (1) and (2) are necessary. That question on the inclusiveness of a classifier touches on ethics of equity in the performance of algorithms.

Although the presented approach shows validity in addressing the issue of bias, there are still a number of threads that need to be investigated. Future work in this domain includes validating the approach with a focus on a different minority class, validating the approach with a focus on a different classification problem, and validating the approach with different generalized machine learning algorithms. We will also target improving the classification rate of both Fear and Disgust, since both of these expressions are hard to detect, and would provide further evidence of the impact of this methodology. We hope that this work will contribute to raising the sensitivity to the potential challenges in the performance and bias of classifiers when making inferences about people of different ages and skin colors. There are opportunities for additional research to identify and address these challenges.

V. References

[1] Kotsiantis, S., Kanellopoulos, D. and Pintelas, P. "Handling imbalanced datasets: A review," *GESTS International Transactions on Computer Science and Engineering*, Vol. 30, pp. 25-36, 2006.

[2] Chawla, N.V., Hall, L. O., Bowyer, K. W. and Kegelmeyer, W. P., "SMOTE: Synthetic Minority Oversampling Technique," *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321–357, 2002.

[3] Joshi, M. V., Kumar, V. and Agarwal, R.C. "Evaluating boosting algorithms to classify rare cases: comparison and improvements," In First IEEE International Conference on Data Mining, pp. 257-264, 2001.

[4] Provost, F. and Fawcett, T. "Robust classification for imprecise environments," *Machine Learning*, Vol. 42, pp. 203-231, 2001.

[5] Doerrfeld, B. "20+ Emotion Recognition APIs That Will Leave You Impressed, and Concerned," http://nordicapis.com/20-emotion-recognition-apis-that-will-leave-you-impressed-and-concerned, 2015.

[6] Dhall, A., Ramana Murthy O.V., Goecke, R., Joshi J. and Gedeon, T. "Video and Image based Emotion Recognition Challenges in the Wild: EmotiW 2015," ACM International Conference on Multimodal Interaction (ICMI), 2015.

[7] Brown, L. and Howard A. "Gestural Behavioral Implementation on a Humanoid Robotic Platform for Effective Social Interaction," IEEE Int. Symp. on Robot and Human Interactive Communication (RO-MAN), pp. 471 – 476, 2014.

[8] Metta, G., Sandini, G., Vernon, D. Natale, L. and Nori, F. "The iCub humanoid robot: an open platform for research in embodied cognition," 8th workshop on performance metrics for intelligent systems, pp. 50-56, 2008.

[9] Cloutier, P., Park, H.W., MacCalla, J. and Howard, A. "It's All in the Eyes: Designing Facial Expressions for an Interactive Robot Therapy Coach for Children," 8th Cambridge Workshop on Universal Access and Assistive Technology, Cambridge, UK, 2016.

[10] Simmons R., et al. "GRACE: An Autonomous Robot for the AAAI Robot Challenge," *AI Magazine*, Vol. 24(2), pp. 51-72, 2003.

[11] Sankur, B., Ulukaya, S. and Çeliktutan, O. "A Comparative Study of Face Landmarking Techniques," *EURASIP J. Image and Video Processing*, Vol. 13, 2013.

[12] Littleworth, G., Bartlett, M.S., Salamanca, L.P. and Reilly, J. "Automated Measurement of Children's Facial Expressions during Problem Solving Tasks," IEEE Int. Conference on Automatic Face and Gesture Recognition, pp. 30–35, 2011.

[13] Albu, F., Hagiescu, D., Vladutu, L. and Puica, M. "Neural network approaches for children's emotion recognition in intelligent learning applications," in Proc. of EDULEARN 2015, Barcelona, Spain, pp. 3229-3239, 2015.

[14] Dalrymple, KA, Gomez, J, and Duchaine, B. "The Dartmouth Database of Children's Faces: Acquisition and Validation of a New Face Stimulus Set," *Urgesi C, ed.PLoS ONE. 2013*. Vol. 8(11), 2013.

[15] Khan, R.A., Meyer, A. and Bouakaz, S. "Automatic Affect Analysis: From Children to Adults," International Symposium on Visual Computing, ISVC 2015, pp. 304-313, 2015.

[16] Egger, H.L., Pine, D.S., Nelson, E., et al. "The NIMH Child Emotional Faces Picture Set (NIMH-ChEFS): A new set of children's facial emotion stimuli," *International Journal of Methods in Psychiatric Research,* Vol. 20(3), pp. 145-156, 2011.

[17] Schmidt, A. "Cloud-Based AI for Pervasive Applications," *IEEE Pervasive Computing*, Vol. 15(1), pp. 14-18, 2016.

[18] Barsoum, E., Zhang, C., Canton Ferrer, C. and Zhang, Z. "Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution," ACM International Conference on Multimodal Interaction (ICMI), Tokyo, Japan, 2016.

[19] Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H.J., et. al. "Presentation and validation of the Radboud Faces Database," *Cognition & Emotion,* Vol. 4(8), pp. 1377—1388, 2010.

[20] Romani-Sponchiado, A., Sanvicente-Vieira, B., Mottin, C., Hertzog-Fonini, D., Arteche, A. "Child Emotions Picture Set (CEPS): Development of a database of children's emotional expressions," *Psychology & Neuroscience*, Vol. 8(4), pp. 467-478, 2015.

[21] Gagnon, M., Gosselin P., Hudon-ven der Buhs, I., Larocque, K., Milliard, K. "Children's recognition and discrimination of Fear and Disgust facial expressions," *Journal of Nonverbal Behavior*, Vol. 34(1), pp. 27–42, 2010.

[22] Gosselin, P., Roberge, P. and Lavalle´e, M. C. "The development of the recognition of facial emotional expressions comprised in the human repertoire," *Enfance*, Vol. 4, pp. 379–396, 1995.

[23] Ekman, P. and Friesen, W. Facial action coding system: A technique for the measurement of facial movement. Palo Alto, Ca.: Consulting Psychologists Press, 1978.

[24] Friesen, W. and Ekman, P. EMFACS-7: Emotional Facial Action Coding System. Unpublished manual, University of California, California, 1983.

[25] Matsumoto, D. and Ekman, P. "Facial expression analysis," *Scholarpedia*, Vol. 3(5), pp. 4237, 2008.

[26] Grd, P. "Two-dimensional face image classification for distinguishing children from adults based on anthropometry," Thesis submitted to University of Zagreb, 2015.

[27] Alom M.Z., Piao, M-L., Islam M.S., Kim, N. and Park, J-H. "Optimized Facial Features-based Age Classification," World Academy of Science. Engineering and Technology Conference, Vol. 6, pp. 319–324, 2012.

[28] Joachims, T. "Making Large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning," B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

[29] Crawford, K. "Artificial Intelligence's White Guy Problem," New York Times – Opinion, http://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html, 2016.

[30] Brodny, G., Kołakowska, A., Landowska, A., Szwoch, M., Szwoch, W. and Wróbel, M. "Comparison of selected off-the-shelf solutions for emotion recognition based on facial expressions," 9th Int. Conf. on Human System Interactions (HSI), pp. 397-404, 2016.