



A comparison of humans and machine learning classifiers categorizing emotion from faces with different coverings



Harisu Abdullahi Shehu ^{a,*}, Will N. Browne ^b, Hedwig Eisenbarth ^c

^a School of Engineering and Computer Science, Victoria University of Wellington, New Zealand

^b School of Electrical Engineering and Robotics, Queensland University of Technology, Australia

^c School of Psychology, Victoria University of Wellington, New Zealand

ARTICLE INFO

Article history:

Received 3 December 2021

Received in revised form 26 September 2022

Accepted 3 October 2022

Available online 10 October 2022

Keywords:

Emotion classification

Emotion categorization

Face masks for COVID-19

Facial expression

Sunglasses

ABSTRACT

Partial face coverings such as sunglasses and face masks unintentionally obscure facial expressions, causing a loss of accuracy when humans and computer systems attempt to categorize emotion. With the rise of soft computing techniques interacting with humans, it is important to know not just their accuracy, but also the confusion errors being made—do humans make less random/damaging errors than soft computing? We analyzed the impact of sunglasses and different face masks on the ability to categorize emotional facial expressions in humans and computer systems. Computer systems, represented by VGG19, ResNet50, and InceptionV3 deep learning algorithms, and humans assessed images of people with varying emotional facial expressions and with four different types of coverings, i.e. unmasked, with a mask covering the lower face, a partial mask with transparent mouth window, and with sunglasses. The first contribution of this work is that computer systems were found to be better classifiers (98.48%) than humans (82.72%) for faces without covering (>15% difference). This difference is due to the significantly lower accuracy in categorizing anger, disgust, and fear expressions by humans ($p < .001$). However, the most novel aspect of the work is identifying how soft computing systems make different mistakes to humans on the same data. Humans mainly confuse unclear expressions as neutral emotion, which minimizes affective effects. Conversely, soft techniques often confuse unclear expressions as other emotion categories, which could lead to opposing decisions being made, e.g. a robot categorizing a fearful user as happy. Importantly, the variation in the misclassification can be adjusted by variations in the balance of categories in the training set.

© 2022 Elsevier B.V. All rights reserved.

1. Introduction

Facial expressions convey information regarding human emotion in non-verbal communication. Thus, an accurate categorization of a person's emotional facial expression contributes to understanding their needs, intentions, and potential actions. This is a valuable insight for soft computing applications such as when robots are introduced into shared workspaces. Thus, computational methods are needed to analyze facial expressions, but how do they differ from human classifiers, especially in different circumstances?

Face coverings such as sunglasses and face masks are used to cover the face for different reasons. For instance, people use

sunglasses to either protect the eyes from sunlight or to improve appearance, and people use face masks often to prevent the spread of infectious diseases such as the coronavirus (COVID-19) or for hygienic reasons in hospitals or food preparation. However, these coverings not only cover parts of a face but might also affect social interaction as they obscure parts of facial expressions, which might contain socially relevant information even if the expression does not represent the emotional state of a person. While sunglasses cover an estimated 10%–15% of the face, approximately 60%–65% of the face is covered with face masks — exact numbers are hard to measure due to variance across different people, dependent on the size of the face [1]. More importantly, these coverings obscure different parts of the face, which convey information about emotional expressions to a varying extent [2].

Different studies have shown that humans are far from perfect in accessing the emotional states of people by just looking at their faces [3]. Further, humans are also far from perfect when categorizing emotional labels from facial expressions [4].

The code (and data) in this article has been certified as Reproducible by Code Ocean: (<https://codeocean.com/>). More information on the Reproducibility Badge Initiative is available at <https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals>.

* Corresponding author.

E-mail address: harisushehu@ecs.vu.ac.nz (H.A. Shehu).

Occlusion of the face leads to an even higher decrease in the performance of humans in categorizing emotion from the faces of people. For instance, it has been observed that partial occlusion of the eyes [5] impaired the achieved performance by humans especially for categorizing sad expressions [6–8]. While negative emotions such as sadness and fear are the most difficult to categorize by humans [9,10] compared with positive emotions like happiness [11–14], obscuring the mouth leads to less accurate categorization even for happiness [6].

More specifically, Carbon [1] investigated the impact of face masks on humans categorizing six different categories of emotion (anger, disgust, fear, happy, neutral, and sad) from images of people. They found a significant decrease in performance of their participants across all expressions except for fear and neutral expressions. In addition, emotional expressions such as happy, sad, and anger were repeatedly confused with neutral expressions whereas other emotions like disgust were confused with anger expressions. In accordance with Carbon, several studies confirmed the detrimental effect of face masks on emotion categorization by humans [5,15,16]. However, none of these studies directly compared the impact of face masks on humans with machine learning classifiers (computer systems).

In contrast to humans, computer systems have done well in categorizing emotion from in-distribution faces images (images from the same dataset) of people, achieving an accuracy of over 95% [17]. Nevertheless, the performance of these systems is greatly affected when it comes to categorizing emotion from natural images where the sun is free to shine and people are free to move, changing dramatically the appearance of faces in these images [18]. For instance, a change in the color information of the pixels, e.g. foreground color of the whole or a small region of the face, which are nonlinearly confounded with emotional expressions, can lead to a decrease in the performance of these models by up to 25% [19,20]. Furthermore, face coverings such as sunglasses and face masks affect the performance of computer systems by up to 74% [21]. However, it is unknown to which extent the same type of sunglasses and face masks would affect the performance of humans in direct comparison to computer systems in emotion categorization of the same stimuli.

We aim to analyze the limitations of humans and computer systems in categorizing emotion of covered and uncovered faces, as well as the misclassification cost of the computer systems.

The contributions of this work are through the following tasks:

- Compare the impact of sunglasses and different types of face masks (i.e. fully covered and recently introduced face masks with a transparent window [22]) on humans' and computer systems' ability to categorize emotional facial expressions. This direct comparison using the same stimuli across different emotion categories and different standardized face coverings addresses open questions about the different importance of facial features for both types of observers.
- Analyze the misclassification cost for the computer systems. This has practical importance as when positive expressions are misclassified as negative emotional expressions, artificial intelligence systems might refrain from interacting with a person. Conversely, the systems might continuously try to interact with a person if an expression such as anger or sadness were misclassified as happiness, which could lead to potential problems in the interaction of humans with these systems.
- Introduce an approach to minimize the misclassification costs for the computer systems while maintaining an equivalent (or minimally reduced) accuracy.

The main reason for analyzing the misclassification cost for the computer systems is because classifier induction often assumes that the best distribution of classes within the data is an equal distribution, i.e. where there is no class imbalance [23–26]. However, research has shown that this assumption is often not true and might lead to a model that is very costly when it misclassifies an instance [27–29]. This might also be true for emotion classification [21], e.g. when the fully covering mask was used to cover the face, emotion categories like happiness were classified badly by the computer systems due to the majority of the happy expression images being misclassified as anger expression. Thus, a secondary objective is to investigate how data distribution affects misclassification for computer systems.

This is important because it is not the error rate that matters. In most cases, what matters most is "where" the misclassification occurs, i.e. the negative effect that may be caused by computer systems misclassifying human facial expressions. For these reasons, the cost of misclassification needs to be assessed for computer systems trained with data that is oversampled based on an expression that is considered to be of less threat. This is anticipated to change the behavior of the computer systems such that if they misclassify unclear expressions to expressions, it will cause the least harm. Therefore, as additional analysis, we will introduce an approach to minimize the misclassification costs for the computer systems.

Considering that the face coverings balance ecological validity with standardization, by obscuring facial expressions as it happens with varieties of face masks and sunglasses, while still being standardized across the individuals and images in the dataset. Understanding these limitations and regulating the cost of misclassifying faces with these coverings can inform the production of artificial emotionally intelligent systems [30] when it comes to real-world situations.

The rest of the paper is organized as: Section 2 describes the machine learning classifiers used as examples of artificial emotion classification systems and provides details on how the classifiers were set up in each experiment. The section also describes the human sample, the survey set up, the tasks needed to be performed by the participants during the survey, and the analysis procedure used. Section 3 presents the obtained results, along with the statistical analysis carried out. Section 4 provides a further discussion on the obtained results, and finally, Section 5 concludes the paper.

2. Method

This study involved collecting and analyzing data from humans and computer systems to compare their accuracy and confidence in categorizing emotion from covered and uncovered faces.

2.1. Human sample

Ethical approval for this study has been obtained from the Human Ethics Committee (HEC) of Victoria University of Wellington and the study was preregistered on OSF (<https://osf.io/mgx9p>). Participants gave informed consent for participation in the study and could enter a raffle/draw to win a \$50 voucher for compensation.

All participants were recruited online via SurveyCircle and social media sites such as Facebook, Twitter, Instagram, etc. through snowball sampling. A power calculation based on the effect size obtained from categorizing the "afraid" class from Eisenbarth et al.'s [31] study was performed to estimate the required sample size, which was found to be 52. However, the study aimed to

recruit at least 100 participants to prepare for potential missing data.

Of 154 people who started participating in the survey, 82 participants completed the full study. Of those 82 participants, 43 were male, 37 were female, and 2 either did not specify their gender or identified themselves to be from other groups. The age of the participants ranged from 19 to 70, with a mean of 27.88 (SD = 7.53). 28 of the participants self-identified as white, 11 as Black or African American, 24 as Asian, 4 as Arab, 3 as multicultural, and the remaining 12 either did not answer the question or identified themselves to be from other groups.

Data were collected via SoSciSurvey (<https://www.soscisurvey.de/>) between the 21st of December 2020 to the 21st of February 2021 during the COVID-19 pandemic.

2.2. Computer systems sample

The computer systems sample was represented by the artificial emotion classification systems. Three of the most commonly used deep learning (i.e. the Visual Geometry Group (VGG) [32], the Residual Neural Network (ResNet) [33], and the Inception (Inception) [34]) models were used as examples of a machine learning algorithms (computer systems).

The Visual Geometry Group (VGG) model is one of the most popular convolutional neural network architectures. The network made significant improvements over AlexNet [35] by replacing the large kernel-sized filters with several 3×3 kernel-sized filters. There are two types of VGG models; the VGG16 and the VGG19 model. VGG19 [32] was chosen to be used as it has more weight layers (i.e. 19) compared to its pair (VGG16) with 16 weight layers. Another reason for selecting VGG19 is because it is a well-studied standard model that has achieved high performance [36,37].

The residual neural network (ResNet) is a powerful neural network that utilizes skip connections, or jumps over certain layers, to avoid the problem of vanishing gradient and mitigates the degradation problem, which causes poor learning for deep networks. There are different types of ResNet such as ResNet18, ResNet34, ResNet50 etc. ResNet50 [33] was chosen as it improves the efficiency of the network with more layers while minimizing the percentage classification error.

InceptionV3 is a widely used convolutional neural network architecture that is made up of symmetric and asymmetric building blocks. The model applied batch normalization to activation functions throughout the model to propagate label information across the network, which stabilizes the learning process. There are different types of the Inception network such as Inception version 1 (InceptionV1), Inception version 2 (InceptionV2), Inception version 3 (InceptionV3), etc. InceptionV3 [34] was chosen because it provides several improvements, e.g. use of auxiliary classifier and factorization of convolutions, to its previous versions. As a result, the network is faster and lighter compared to versions 1 and 2.

The models were set up using the Keras API [38] to use the same number of layers as originally introduced in each corresponding paper [32–34]. The models were all set up to run for 200 epochs, starting with an initial learning rate of 0.001. The learning rate was reduced by 10% after 80, 100, 120, 160 epochs, and by 5% after 180 epochs to avoid overfitting¹ the data. Note that the computer systems were trained with images from the last-half frames of the CK+ dataset based on the technique developed by Shehu et al. [39].

¹ Overfitting occurs when a model learns details, including noise in the training data to such an extent that it has a negative impact on the performance of the model on new data. As a result, the model performs too well on the training data, but poorly on the test data.

The accuracy of the deep models was obtained by dividing the number of correctly predicted test images by the total number of presented test images, which in this case is 35 for each type of covering. In deep models, the prediction of each image returns the probability of associating the image to a particular emotion category where the summation of all the probabilities is equal to 1 ($\sum_{i=1}^n P_i = 1$). The predicted class is the class with the highest probability ($\max_{x \in [1, \dots, n]} P(x)$ where n is the number of emotion categories). These probabilities are usually considered as the confidence of the model.

The three computer systems (VGG19, ResNet50, and InceptionV3) algorithms were included to test generalizability. However, detailed results of only the VGG19 have been provided as it has achieved higher performance compared to ResNet50 and the InceptionV3 models on partially covered images of the CK+ dataset.

Due to the stochastic nature of processes and the non-deterministic nature of the deep models, the model was run 30 times and the results presented are an average obtained from the 30 runs. Thus, the study compared the survey results obtained from 82 human participants with the 30 computer systems.

2.3. Stimulus material

Seven categories of emotion, i.e. the six basic emotions (*anger*, *disgust*, *fear*, *happy*, *sad*, and *surprise*) defined by Ekman [40], as well as the neutral expression were included. The experiment consisted of four blocks presenting facial expressions unmasked, masked, partially masked, and with sunglasses. The unmasked block consisted of images with nothing added to the face. Five images were used for each of the seven categories, i.e. a total of 35 images are used. The images used mainly comprised images of different people such that participants could not classify the facial expression from a previous image seen. Similarly, the masked block consisted of 35 emotional images of people with face masks added to the images. The partially masked block consisted of 35 emotional images of people with a partial mask, i.e. a face mask with a transparent window added to the images and finally, the sunglasses block consisted of 35 images of people with sunglasses added to the images. All the images used in the experiment are emotional images from the CK+ dataset as images from the dataset have recently been used in numerous and wide ranging studies [41–45]. However, only one section (unmasked) consisted of original images from the CK+ dataset. The remaining three blocks (mask, partial mask, and sunglasses) consisted of simulated images with different kinds of face masks, as well as sunglasses. Face masks and sunglasses were added to the images using a strategy that was implemented in our previous research [21]. Fig. 1(a) shows sample original and simulated emotional images of the six basic emotions, plus the neutral expression used here.

This research used a total of 3018 images from the last-half frame images of the CK+ dataset to train the computer systems based on the technique developed by Shehu et al. [17,39]. The test images were randomly selected from the last-half frame images and not used in system training. Thus, the test images are a combination of images from mid to peak or highest emotional intensity frames.

2.4. Descriptive questions

Human participants were asked to answer a set of questions such as their age, gender, and ethnicity. Furthermore, we asked the participants three additional questions to understand the extent of their prior experience with people wearing masks.



(a)



What emotion do you think this person is showing?



Anger

Fear

Neutral

Surprise

Disgust

Happy

Sad

How confident are you that the image belongs to the category you have selected?



Next

(b)

Fig. 1. Sample images of faces: (a) unmasked (first row), masked (second row), partial mask (third row), and sunglasses (last row) from the CK+ dataset. Left to right: anger, disgust, fear, happy, neutral, sad, and surprise expressions. [39] (b) example trial for human participant.

The three questions asked to the participants were; “(1) For how long did you interact with people wearing masks on a daily basis during the past 3 months?”, “(2) On average, for how much time per day (in percentage) did you interact with people wearing masks during the past 3 months?”, and “(3) Of the people you saw on a day-to-day basis during the past 3 months, approximately [what] percentage [was] wearing masks on average?”. All questions were asked prior to taking the survey and participants answered using a slider ranging from 0 (“not at all”) to 100

(“very often”). A mean exposure score (Qmean) was computed, averaging the three questions ($M = 36.15$, $SD = 27.62$).

2.5. Emotion categorization task

Fig. 1(b) shows a sample trial. For each trial, participants were required to choose the emotion category that the image most closely related to, as well as their confidence in choosing the category of the emotion. There was a total of 140 trials across the 7 categories. All emotion categorization questions were required

to be answered and participants were not taken to the next page unless the emotion category and the confidence rating were made (see Supplementary Materials).

The data collected, as well as the stimulus set used, can be accessed via OSF (<https://osf.io/mgx9p/>).

2.6. Analysis

Group and stimulus effects on accuracy and confidence ratings were tested using a $2 \times 4 \times 7$ repeated-measures ANOVA with the factor Emotion category with 7 levels (anger, disgust, fear, happy, neutral, sad, and surprise), the factor group (computer systems vs humans) and the factor covering (unmasked, masked, partial mask, and sunglasses). Post hoc tests included 2×4 repeated measures ANOVAs and t-tests.

For all multiple tests such as post hoc tests after the initial overall ANOVAs, the alpha level was adjusted using Bonferroni corrections [46].

3. Results

This section presents the accuracy ratings obtained from humans compared with computer systems (i.e. VGG19, ResNet50, and InceptionV3) in categorizing emotion from covered and uncovered faces.

3.1. Emotion categorization

Initially, a 3-way repeated-measures ANOVA ($2 \times 4 \times 7$) was performed with the between-subject factor Group (i.e. humans vs computer systems), the within-subject factors Covering (i.e. unmasked, mask, partial mask, and sunglasses), and Emotion (i.e. anger, disgust, fear, happy, neutral, sad, and surprise) to analyze the achieved accuracy. We found significant main effects for Group, $F(1, 3080) = 80.25, p < .001, \eta^2 = 0.01$, Covering, $F(3, 3080) = 40.44, p < .001, \eta^2 = 0.09$ and Emotion, $F(6, 3080) = 81.87, p < .001, \eta^2 = 0.02$. All interactions between the factors were significant (see Table 1a), as well as the three-way interaction of Group \times Covering \times Emotion $F(18, 3080) = 28.11, p < .001, \eta^2 = 0.09$.

Thereafter, the alpha level was adjusted using Bonferroni correction ($\alpha = 0.0083$) before post hoc tests. A significant interaction between the emotion categories (i.e. anger, disgust, fear, happy, neutral, sad, and surprise) was found for humans, $F(6, 567) = 33.18, p < .001, \eta^2 = 0.26$ and computer systems, $F(6, 203) = 33.14, p < .001, \eta^2 = 0.50$ for the unmasked emotion (see Fig. 2(a)).

Fig. 2(a) represents the violin plots for categorization accuracy of unmasked images of people by both humans and computer systems. As can be seen from Fig. 2(a), the computer systems classify almost all the classes correctly, achieving an overall accuracy of 98.48% compared to humans who achieved an overall accuracy of 82.72%. This decrease in the overall performance of humans is based on significantly lower accuracy in categorizing anger, disgust, and fear expressions, with fear contributing the most (Error rate $> 30\%$) to the significantly worst performance (see Table 1b) in humans. However, Bonferroni corrected post hoc t-tests showed no significant difference in categorization accuracies achieved by the computer systems across emotion categories except for the neutral expression (see Table 1c), which was categorized with a significantly lower accuracy compared to other emotion categories.

The classification accuracy achieved by the computer systems significantly decreased from an average accuracy of 98.48% when there was no mask to 24.0% when there were masks on the face (see Fig. 2(b)). Not only did the average accuracy decrease, but

also the accuracy achieved from categorizing each class decreased significantly ($> 68\%$), except for anger (less than 16%). Although the overall accuracy achieved by humans for masks on the face also decreases in each category (from when there was no mask), the decrease is not large ($< 26\%$) compared to the decrease seen in the computer systems' accuracy (up to 74.48%). Also, post hoc tests showed that differences between emotion categories were significant for both humans, $F(6, 567) = 64.85, p < .001, \eta^2 = 0.41$ and computer systems, $F(6, 203) = 125.79, p < .001, \eta^2 = 0.79$ ($\alpha = 0.0083$) for the masked faces. Post hoc two-sample t-tests for the accuracies in each category obtained from humans and the computer systems showed significant differences ($p < .001$) for most of the emotion categories at alpha = 0.0014 (see Table 1d and e).

The violin plots presented in Fig. 2(c) show accuracies obtained from categorizing images by humans and computer systems when a partial mask (mask with transparent window) was used to cover the face. In contrast to the use of a fully covering mask, the classification accuracy obtained by computer systems saw a great increase in categorizing happiness and surprise expressions with partial masks (0% vs 66.67% and 20% vs 73.33%). Also, while the accuracy achieved by humans had also increased across all the emotion categories, the largest increase was seen for happy, disgust, sad, and surprise expressions (all $> 19\%$ increase). Post hoc tests revealed significance differences for both humans, $F(6, 567) = 64.01, p < .001, \eta^2 = 0.40$ and computer systems, $F(6, 203) = 235.28, p < .001, \eta^2 = 0.87$ ($\alpha = 0.0083$). Post hoc two-sample unpaired t-tests showed a significant differences ($<.001$) for most emotion categories (see Table 2a and b).

The violin plots presented in Fig. 2(d) show the accuracy obtained from categorizing faces with sunglasses by humans and computer systems. As can be seen, the accuracy achieved by the computer systems decreased most for sad and surprise expressions with sunglasses (all $> 27\%$). Notably, the accuracy achieved by humans in categorizing happiness remains the same both with or without sunglasses. This shows that putting on sunglasses does not reduce humans' happiness detection. As expected, there was a large decrease ($> 16\%$) in the accuracy for fear expressions when the eyes were covered. Post hoc tests showed that these differences between emotion categories were significant for both humans, $F(6, 567) = 74.15, p < .001, \eta^2 = 0.44$ and computer systems, $F(6, 203) = 13.85, p < .001, \eta^2 = 0.29$ ($\alpha = 0.0083$) for faces with sunglasses. Post hoc two-sample unpaired t-test showed significant differences ($<.001$) for most of the emotion categories (see Table 2c and d).

We also tested for differences within each emotion category across different types of covering (i.e. unmasked, mask, partial mask, and sunglasses). A significant main effect was found for all emotion categories (all $p < .001$), except for the neutral expression, $F(3, 324) = 0.037, p = .99, \eta^2 = 0.003$ (see Fig. 3(a)). Although, there was a significant, $F(3, 324) = 5.36, p < .001$ ($p = 0.00129$), $\eta^2 = 0.05$ main effect for humans categorizing the fear expression comparing the different types of coverings (see Fig. 3(b)), the effect size was very small ($\eta^2 = 0.05$) compared with other emotion categories (all $\eta^2 > 0.2$) like anger (see Fig. 3(c)), etc. that have a larger effect.

Thus, humans might be equally worse (see Fig. 3(b)) in categorizing the fear emotion category with or without covering on the face as the effect size is not as large as it is in other expressions (all $\eta^2 > 0.2$ except neutral expression $\eta^2 = 0.003$).

Based on the three additional questions asked about exposure to face masked people, Qmean was added as a factor to a linear regression and then investigate if people's experience with masked individuals might help them achieve higher accuracy in

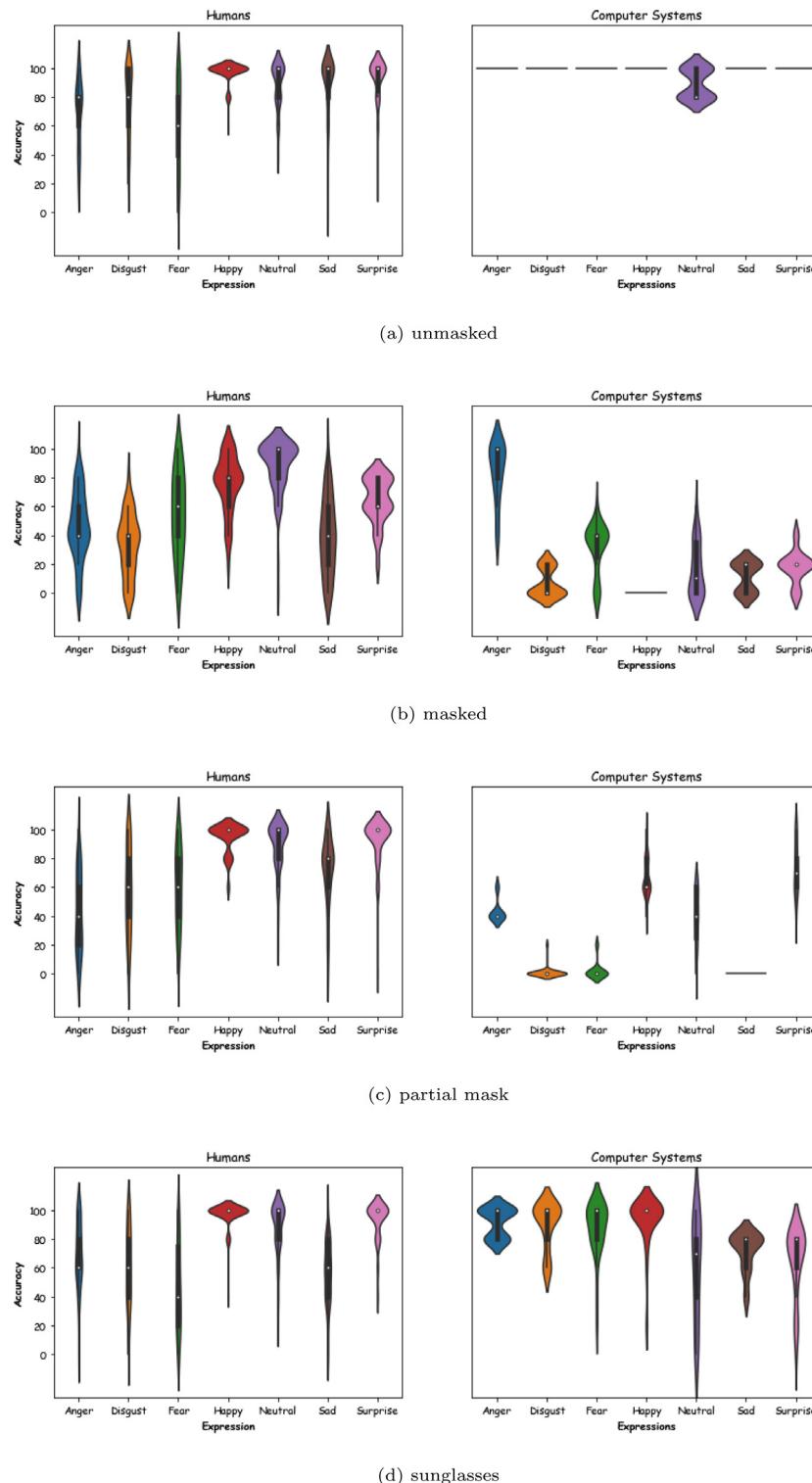


Fig. 2. Violin plots showing accuracy obtained from categorizing (a) unmasked, (b) masked, (c) partial mask, and (d) sunglasses emotions by humans and computer systems (VGG19).

categorizing mask emotion images. The results (see Table 2e) showed significant main effects for the factors Covering, $F(3, 6) = 122.19, p < .001, \eta^2 = 0.09$ and Qmean, $F(6, 70) = 3.73, p < .001, \eta^2 = 0.07$, where Qmean represents the average of the three questions (higher accuracy related to higher Qmean mean more exposure). However, there was no significant interaction between Covering and Qmean, $F(6, 210) = 0.62, p = 1.00, \eta^2 = 0.03$.

3.2. Misclassification analysis

As an additional analysis, this research investigated the types of misclassification made by humans and computer systems.

The results (see Section 3.1) showed that humans mainly misclassify expressions with full covering (see Fig. 4(a)) and therefore, unclear expressions as neutral. However, the misclassifications varied for the computer systems. For instance,

Table 1

Statistical test results presenting (a) ANOVA obtained from comparing humans and computer systems based on their achieved accuracy from categorizing people's emotions with different types of coverings. (b) Post hoc *t*-test results showing the *p*-value and Cohen's *d* effect size obtained from comparing unmasked emotion categories by humans and (c) computer systems. (d) Post hoc *t*-test results showing the *p*-value and Cohen's *d* effect size obtained from comparing masked emotion categories by humans and (e) computer systems. Note that the α level in all the post hoc comparisons has been adjusted to 0.0014 using Bonferroni correction.

(a) ANOVA comparing humans and computer systems							
	Sum square	df	F	<i>p</i>	Eta squared		
Intercept	1.76E+05	1.0	438.05	<.001	0.08		
C(Group)	3.23E+04	1.0	80.25	<.001	0.01		
C(Covering)	4.88E+04	3.0	40.44	<.001	0.09		
C(Emotion)	1.98E+05	6.0	81.87	<.001	0.02		
C(Group): C(Covering)	1.78E+04	3.0	14.73	<.001	0.09		
C(Group): C(Emotion)	1.94E+05	6.0	80.44	<.001	0.01		
C(Covering): C(Emotion)	1.56E+05	18.0	21.55	<.001	0.07		
C(Group): C(Covering):C(Emotion)	2.03E+05	18.0	28.11	<.001	0.09		
Residual	1.24E+06	3080.0	—	—	—		
(b) Unmasked by humans							
	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise
Anger	—	0.123 (-0.4)	0.341 (0.25)	<.001 (-1.63)	<.001 (-1.00)	<.001 (-0.96)	<.001 (-1.20)
Disgust	0.123 (-0.40)	—	0.025 (0.60)	<.001 (-1.08)	0.049 (-0.52)	0.048 (-0.52)	0.008 (-0.72)
Fear	0.341 (0.25)	0.025 (0.60)	—	<.001 (-1.59)	<.001 (-1.11)	1.00 (-1.08)	1.00 (-1.27)
Happy	<.001 (-1.63)	<.001 (-1.08)	<.001 (-1.59)	—	0.008 (0.71)	0.056 (0.51)	0.113 (0.42)
Neutral	<.001 (-1.00)	0.049 (-0.52)	<.001 (-1.11)	0.008 (0.71)	—	0.832 (-0.06)	0.354 (-0.24)
Sad	<.001 (-0.96)	0.048 (-0.52)	<.001 (-1.08)	0.056 (0.51)	0.832 (-0.06)	—	0.558 (-0.15)
Surprise	<.001 (-1.20)	0.008 (-0.72)	<.001 (-1.27)	0.113 (0.42)	0.354 (-0.24)	0.558 (-0.15)	—
(c) Unmasked by computer systems							
	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise
Anger	—	1.00 (0)	1.00 (0)	1.00 (0)	<.001 (1.51)	1.00 (0)	1.00 (0)
Disgust	1.00 (0)	—	1.00 (0)	1.00 (0)	<.001 (1.51)	1.00 (0)	1.00 (0)
Fear	1.00 (0)	1.00 (0)	—	1.00 (0)	<.001 (1.51)	1.00 (0)	1.00 (0)
Happy	1.00 (0)	1.00 (0)	1.00 (0)	—	<.001 (1.51)	1.00 (0)	1.00 (0)
Neutral	<.001 (1.51)	<.001 (1.51)	<.001 (1.51)	<.001 (1.51)	—	<.001 (1.51)	<.001 (1.51)
Sad	1.00 (0)	1.00 (0)	1.00 (0)	1.00 (0)	<.001 (1.51)	—	1.00 (0)
Surprise	1.00 (0)	1.00 (0)	1.00 (0)	1.00 (0)	<.001 (1.51)	1.00 (0)	—
(d) Masked by humans							
	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise
Anger	—	0.027 (0.59)	0.230 (-0.32)	<.001 (-1.36)	<.001 (-2.00)	0.276 (0.29)	<.001 (-0.91)
Disgust	0.027 (0.59)	—	0.002 (-0.85)	<.001 (-2.07)	<.001 (-2.78)	0.327 (-0.26)	<.001 (-1.67)
Fear	0.230 (-0.32)	0.002 (-0.85)	—	0.002 (-0.85)	<.001 (-1.40)	0.036 (0.56)	0.106 (0.43)
Happy	<.001 (-1.36)	<.001 (-2.07)	0.002 (-0.85)	—	0.014 (-0.66)	<.001 (1.58)	0.020 (0.62)
Neutral	<.001 (-2.00)	<.001 (-2.78)	<.001 (-1.40)	0.014 (-0.66)	—	<.001 (2.18)	<.001 (1.38)
Sad	0.276 (0.29)	0.327 (-0.26)	0.036 (0.56)	<.001 (1.58)	<.001 (2.18)	—	<.001 (-1.17)
Surprise	<.001 (-0.91)	<.001 (-1.67)	0.106 (0.43)	0.020 (0.62)	<.001 (1.38)	<.001 (-1.17)	—
(e) Masked by computer systems							
	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise
Anger	—	<.001 (5.03)	<.001 (2.90)	<.001 (6.05)	<.001 (3.61)	<.001 (4.69)	<.001 (4.19)
Disgust	<.001 (5.03)	—	<.001 (-1.80)	<.001 (1.00)	0.017 (-0.64)	0.072 (-0.48)	<.001 (-1.12)
Fear	<.001 (2.90)	<.001 (-1.80)	—	<.001 (2.63)	0.002 (0.87)	<.001 (1.45)	<.001 (0.94)
Happy	<.001 (6.05)	<.001 (1.00)	<.001 (2.63)	—	<.001 (-1.24)	<.001 (-1.62)	<.001 (-2.36)
Neutral	<.001 (3.61)	0.017 (-0.64)	0.002 (0.87)	<.001 (-1.24)	—	0.230 (0.32)	0.605 (-0.13)
Sad	<.001 (4.69)	0.072 (-0.48)	<.001 (1.45)	<.001 (-1.62)	0.230 (0.32)	—	0.017 (-0.64)
Surprise	<.001 (4.19)	<.001 (-1.12)	<.001 (0.94)	<.001 (-2.36)	0.605 (-0.13)	0.017 (-0.64)	—

the computer systems mainly classified full face mask covered expressions as anger, while partially covered facial expressions were mainly classified as happiness or surprise (see Fig. 4(d)).

As such, misclassifications can result in costly negative outcomes and complications for humans interacting with computer systems. Therefore, this work aimed to regulate the cost of misclassification in these computer systems. To make them resemble human classification behavior, we aimed to alter the training methods to also lead to misclassifying unclear expressions as neutral.

A possible explanation of why humans mainly misclassify unclear emotions as neutral is because it is the emotion that they are exposed to predominantly on a day-to-day basis. In other words, humans would classify unclear expressions as neutral because it is the expression they see on people's faces most of the time. Therefore, by oversampling the neutral expression images used to train the computer systems, we anticipated that the models

would behave more similar to humans as they see more neutral expressions compared to other emotion categories.

3.2.1. Method

Since classifying any emotion category as a neutral expression is considered less costly (see Section 3.2) compared to when they are classified as other emotional categories such as anger, happiness, sadness, etc., the misclassification cost of the model was evaluated without taking the samples that were misclassified as neutral into account. In particular, the quality of the classifier is not measured based on total misclassification but rather using Eq. (1) below:

$$\text{cost} = \frac{\text{Err} - \text{Err}_{\text{neutral}}}{\text{No. test samples}} \quad (1)$$

where

Table 2

Statistical test results presenting (a) Post hoc test (*t*-test) results showing the *p*-value and Cohen's *d* effect size obtained from comparing partial mask emotion categories by humans and (b) computer systems. (c) Post hoc test (*t*-test) results showing the *p*-value (and Cohen's *d* effect size) obtained from comparing sunglasses emotion categories by humans and (d) computer systems. (e) Regression results from comparing humans based on their understanding of masked people and their achieved accuracy from categorizing people's emotions with different coverings. Note that the α level in all the post hoc comparisons has been adjusted to 0.0014 using Bonferroni correction.

(a) Partial mask by humans							
	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise
Anger	–	0.051 (−0.52)	0.015 (−0.65)	<.001 (−2.49)	<.001 (−2.02)	<.001 (−1.13)	<.001 (−2.24)
Disgust	0.051 (−0.52)	–	0.695 (−0.10)	<.001 (−1.66)	<.001 (−1.29)	0.047 (−0.53)	<.001 (−1.49)
Fear	0.015 (−0.65)	0.695 (−0.10)	–	<.001 (−1.66)	<.001 (−1.25)	0.096 (−0.44)	<.001 (−1.46)
Happy	<.001 (−2.49)	<.001 (−1.66)	<.001 (−1.66)	–	0.112 (0.42)	<.001 (1.26)	0.621 (0.13)
Neutral	<.001 (−2.02)	<.001 (−1.29)	<.001 (−1.25)	0.112 (0.42)	–	0.002 (0.83)	0.330 (−0.26)
Sad	<.001 (−1.13)	0.047 (−0.53)	0.096 (−0.44)	<.001 (1.26)	0.002 (0.83)	–	<.001 (−1.06)
Surprise	<.001 (−2.24)	<.001 (−1.49)	<.001 (−1.46)	0.621 (0.13)	0.330 (−0.26)	<.001 (−1.06)	–
(b) Partial mask by computer systems							
	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise
Anger	–	<.001 (7.29)	<.001 (6.11)	<.001 (−2.35)	0.566 (0.15)	<.001 (8.22)	<.001 (−2.16)
Disgust	<.001 (7.29)	–	0.310 (−0.27)	<.001 (−7.49)	<.001 (−3.30)	0.319 (0.26)	<.001 (−5.56)
Fear	<.001 (6.11)	0.310 (−0.27)	–	<.001 (−6.85)	<.001 (−3.07)	0.078 (0.47)	<.001 (−5.28)
Happy	<.001 (−2.35)	<.001 (−7.49)	<.001 (−6.85)	–	<.001 (1.72)	<.001 (7.91)	0.104 (−0.43)
Neutral	0.566 (0.15)	<.001 (−3.30)	<.001 (−3.07)	<.001 (1.72)	–	<.001 (3.43)	<.001 (−1.82)
Sad	<.001 (8.22)	0.319 (0.26)	0.078 (0.47)	<.001 (7.91)	<.001 (3.43)	–	<.001 (−5.72)
Surprise	<.001 (−2.16)	<.001 (−5.56)	<.001 (−5.28)	0.104 (−0.43)	<.001 (−1.82)	<.001 (−5.72)	–
(c) Sunglasses by humans							
	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise
Anger	–	0.274 (0.29)	0.002 (0.86)	<.001 (−1.64)	<.001 (−0.95)	0.024 (0.60)	<.001 (−1.30)
Disgust	0.274 (0.29)	–	0.031 (0.57)	<.001 (−1.87)	<.001 (−1.21)	0.292 (0.28)	<.001 (−1.55)
Fear	0.002 (0.86)	0.031 (0.57)	–	<.001 (−2.34)	<.001 (−1.73)	0.169 (−0.36)	<.001 (−2.05)
Happy	<.001 (−1.64)	<.001 (−1.87)	<.001 (−2.34)	–	0.012 (0.68)	<.001 (2.57)	0.158 (0.37)
Neutral	<.001 (−0.95)	<.001 (−1.21)	<.001 (−1.73)	0.012 (0.68)	–	<.001 (1.68)	0.207 (−0.33)
Sad	0.024 (0.60)	0.292 (0.28)	0.169 (−0.36)	<.001 (2.57)	<.001 (1.68)	–	<.001 (−2.12)
Surprise	<.001 (−1.30)	<.001 (−1.55)	<.001 (−2.05)	0.158 (0.37)	0.207 (−0.33)	<.001 (−2.12)	–
(d) Sunglasses by computer systems							
	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise
Anger	–	0.570 (0.15)	0.244 (0.31)	0.576 (−0.15)	<.001 (1.27)	<.001 (1.72)	<.001 (1.49)
Disgust	0.570 (0.15)	–	0.566 (0.15)	0.354 (−0.24)	<.001 (1.12)	<.001 (1.22)	<.001 (1.24)
Fear	0.244 (0.31)	0.566 (0.15)	–	0.158 (−0.38)	<.001 (0.99)	<.001 (0.94)	<.001 (1.05)
Happy	0.576 (−0.15)	0.354 (−0.24)	0.158 (−0.38)	–	<.001 (1.27)	<.001 (1.47)	<.001 (1.42)
Neutral	<.001 (1.27)	<.001 (1.12)	<.001 (0.99)	<.001 (1.27)	–	0.084 (−0.46)	0.550 (−0.16)
Sad	<.001 (1.72)	<.001 (1.22)	<.001 (0.94)	<.001 (1.47)	0.084 (−0.46)	–	0.156 (0.38)
Surprise	<.001 (1.49)	<.001 (1.24)	<.001 (1.05)	<.001 (1.42)	0.550 (−0.16)	0.156 (0.38)	–
(e) ANOVA for human sample including Qmean.							
	Sum Square	df	F	p	Eta Squared		
C(Emotion)	5.28E+05	6.0	172.42	<.001	0.27		
C(Covering)	1.87E+05	3.0	122.19	<.001	0.09		
C(Qmean)	1.33+E05	70.0	3.73	<.001	0.07		
C(1/ID)	5.13+E04	81.0	1.24	0.076	0.03		
C(Covering):C(Qmean)	6.62E+04	210.0	0.62	1.00	0.03		
Residual	1.02+E06	1995.0	–	–	–		

- *Err* represents the misclassified samples or the total misclassification
- *Err_neutral* represents samples misclassified as neutral
- No. *test samples* represents the total number of test sample images

3.2.2. Results

In Table 3, the column *oversampling ratio* represents the oversampling ratio of the category neutral against all other emotion categories. For instance, the value “2:2” means no oversampling whereas the ratio “4:2” means there were twice as much neutral expression images in the training set compared to images from all other emotion categories. Different ratios were chosen based on the number of available images in the CK+ dataset so as to find the best ratio that would give a much lower misclassification cost while maintaining good accuracy.

The results here are presented with the three computer systems (i.e. VGG19, ResNet50, and InceptionV3) to test generalizability. However, a detailed explanation of only the VGG19 model

is provided as the model achieved higher performance compared with the other models (see Table 3).

As can be seen from Table 3, the average misclassification cost obtained in each of the oversampled cases is less than the average cost without oversampling. Thus, oversampling reduced the cost of misclassification, with the least misclassification cost for the models trained with oversampled data at a 5:2 ratio.

Figs. 4(a), 4(b), and 4(c) present the confusion matrices,² obtained from classifying mask, partial mask, and sunglasses images by humans, Figs. 4(d), 4(e), and 4(f) present the confusion matrices obtained from classifying mask, partial mask, and sunglasses images when no oversampling was applied, and Figs. 4(g), 4(h), and 4(i) present the confusion matrices obtained from classifying

² A confusion matrix is a table to visualize the performance of an algorithm. In a confusion matrix, on the basis whereby each row represents the number of instances in a predicted class, the columns will represent the total number of instances in an actual class (or vice-versa).

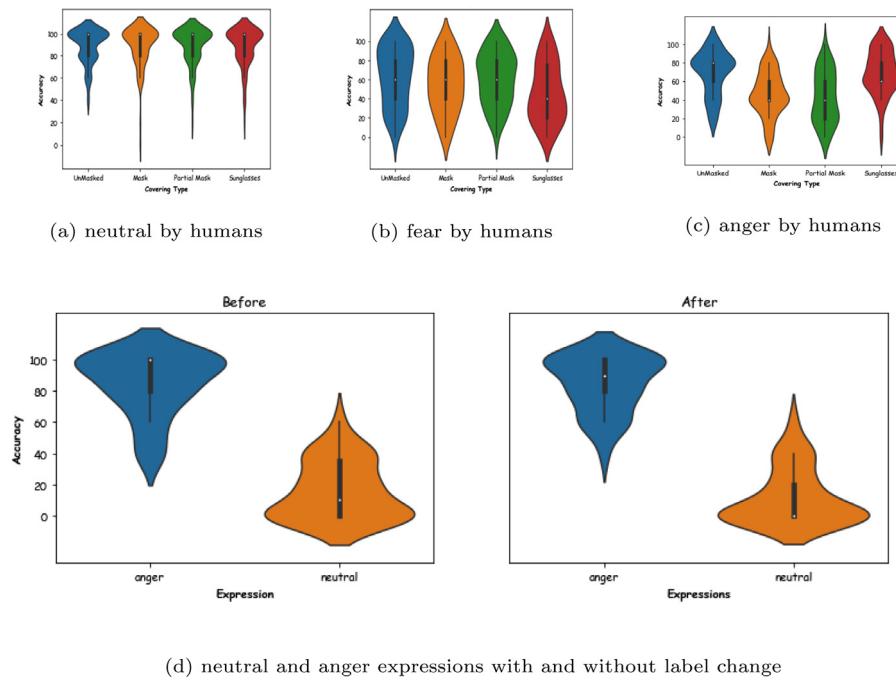


Fig. 3. Violin plot showing the accuracy obtained (a) from classifying neutral expression by humans across four different coverings, (b) from categorizing fear expression by humans across four different coverings, (c) from categorizing anger expression by humans across four different coverings, and (d) by computer systems from categorizing neutral and anger expression before and after the label changed.

Table 3

Table presenting accuracies and the misclassification costs obtained from classifying covered images with and without oversampling neutral expression images.

Oversample ratio (Neutral>All)	Misclassification cost			Avg cost	Accuracy			Avg accuracy	Model
	Mask	Partial mask	Sunglasses		Mask	Partial mask	Sunglasses		
2:2	0.61	0.47	0.18	0.41	0.24	0.33	0.80	0.46	VGG19
3:2	0.40	0.54	0.23	0.39	0.20	0.14	0.68	0.34	
4:2	0.26	0.34	0.17	0.26	0.17	0.23	0.71	0.37	
5:2	0.25	0.39	0.10	0.25	0.25	0.17	0.66	0.36	
2:2	0.41	0.35	0.72	0.49	0.22	0.20	0.23	0.22	ResNet50
3:2	0.33	0.31	0.65	0.43	0.20	0.19	0.22	0.20	
4:2	0.30	0.24	0.56	0.37	0.18	0.15	0.20	0.18	
5:2	0.23	0.19	0.50	0.31	0.16	0.16	0.21	0.18	
2:2	0.69	0.40	0.62	0.57	0.22	0.23	0.34	0.26	InceptionV3
3:2	0.63	0.24	0.64	0.51	0.22	0.19	0.27	0.23	
4:2	0.59	0.21	0.51	0.44	0.23	0.19	0.32	0.25	
5:2	0.54	0.18	0.41	0.38	0.21	0.18	0.38	0.26	

mask, partial mask, and sunglasses images with the best oversampling ratio (i.e. 5:2) applied to the training images (of VGG19). Humans mainly misclassified unclear expressions as neutral when the face was fully mask covered. However, while the misclassification varied for different coverings when no oversampling was applied to computer systems, after oversampling, the majority of the misclassified images were misclassified as neutral. Although the average accuracy achieved by the model without oversampling is larger (>10%) when compared with the average accuracy achieved by the model with the best oversampling ratio (5:2), the difference in the average misclassification cost achieved by the model with the best oversampling ratio was significantly lower (16% lower) than the average misclassification cost of the model with no oversampling, ($t(58) = 30.98, p < .001$).

4. Discussion

The work aimed to compare the performance of humans and computer systems in categorizing emotion from faces of people with sunglasses and face masks, as well as compare the decrease

caused by each covering. The work also introduces an approach that reduces the misclassification cost for computer systems.

In addition to the replication of previous findings that computer systems (with 98.48% accuracy) perform better than humans (with 82.72% accuracy) in categorizing emotions from uncovered/unmasked images in the same dataset [47–50], we found that face masks (both partially covered and fully covering masks) and sunglasses reduced the accuracy of both humans and computer systems in emotion categorization. Consistent with Carbon [1], we also found that humans mainly classify faces with fully covering masks as neutral.

While the face masks impair emotion categorization in both humans and computer systems by a significant amount, emotion categories are affected differently depending on the type of covering used. For instance, the use of a fully covering mask mainly obscured disgust, happiness, and sadness for computer systems whereas anger, disgust, and sadness were mainly more difficult to categorize for humans, which is in line with previous experimental work using stimuli that varied in missing features [6,51]. The partial masks mainly impaired anger categorization in humans, whereas computer systems were mainly impaired through those

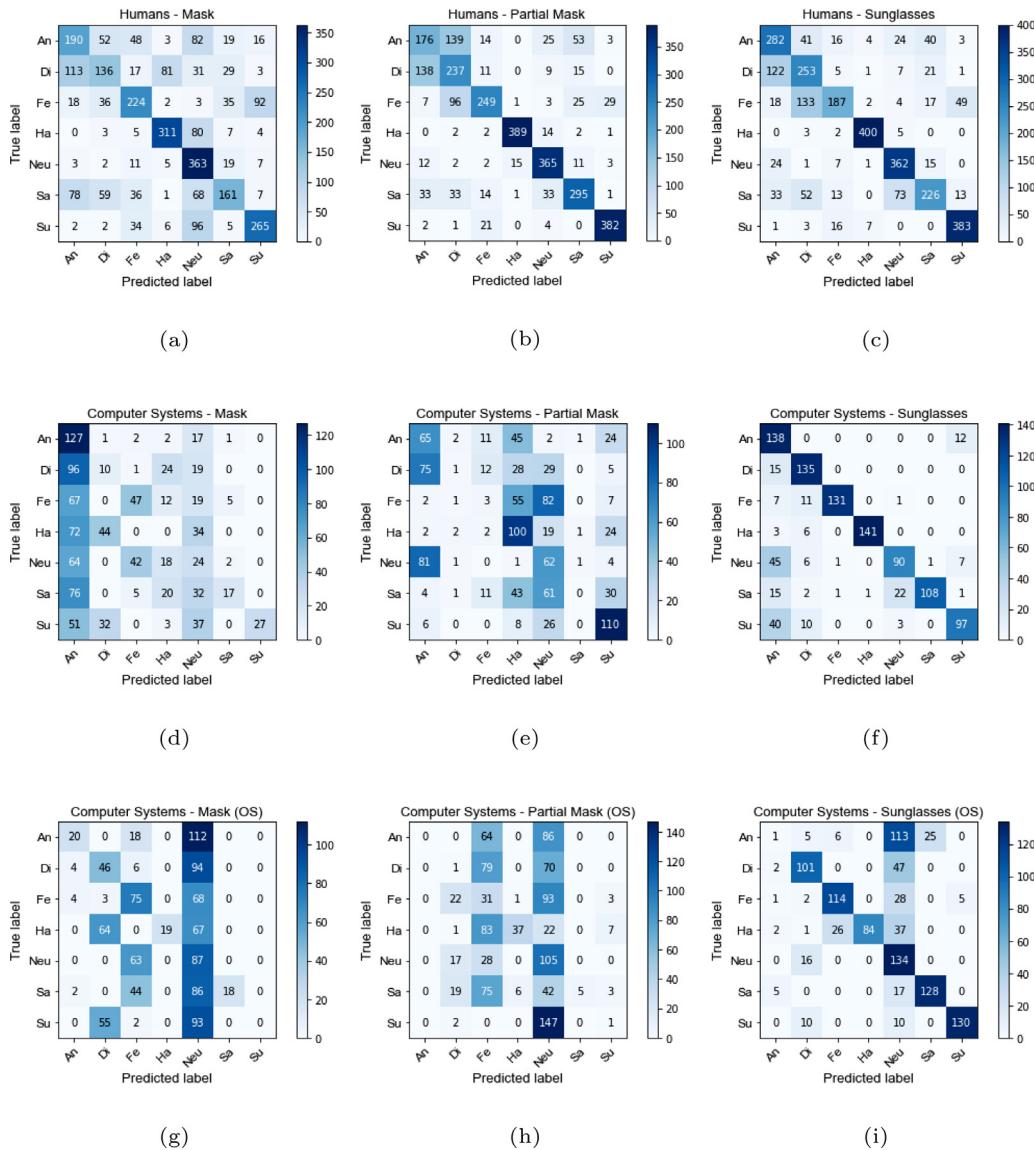


Fig. 4. Confusion matrices obtained from classifying (a) mask, (b) partial mask, and (c) sunglasses images by humans, from classifying (d) mask, (e) partial mask, and (f) sunglasses images by computer systems without oversampling, from classifying (g) mask, (h) partial mask, and (i) sunglasses images by computer systems with oversampling [5:2]. Note that the 82 human participants all classified a sum of 35 images for each covering, i.e. a total of 410 images for each emotion category and 2870 for each covering. Similarly, 30 computer systems (i.e. VGG19) all classified a sum of 35 images for each covering, i.e. a total of 150 images for each emotion category and 1050 for each covering. An = anger, Di = disgust, Fe = fear, Ha = happy, Neu = neutral, Sa = sadness and Su = surprise.

for disgust, fear, and sadness. And while the use of sunglasses mainly reduced fear categorization ability in humans, it was neutral and surprise for computer systems.

Previous work [20,52] has shown the relevance of the mouth for correctly categorizing happiness and surprise in both humans and computer systems. Therefore, since the partial masks allow the mouth to be visible through the transparent window, it was anticipated that happiness and surprise would be categorized more accurately when the partial mask is applied. Indeed, we found a large increase in the classification accuracy of happiness and surprise expression in both computer systems and humans. While there was an increase of 19% (for happy) and 28% (for the surprise) in the human sample, there was an increase of up to 66% for the category happy and up to 55% for surprise expression in computer systems when the partial mask was used.

While we found a significant difference in the accuracy of humans categorizing each emotion category across the different coverings, no significant difference was found for the neutral expression across all coverings. As humans classify unclear

emotions as neutral, they subsequently were highly accurate in categorizing neutral expressions (see Fig. 3(a)).

In contrast to humans, the classification of unclear emotions significantly varies in computer systems. For instance, the computer systems mainly classify full mask covered expressions as anger. However, unclear emotions were mainly classified as happiness or surprise when there was a partial mask on the face. An initial observation of the confusion matrix obtained from categorizing masked emotion images (see Fig. 4(d)) suggested that the computer systems mainly classify unclear emotions as anger expressions just because the anger expression appears first after one-hot encoding³ the labels that were fed to the classifier. Therefore, to test this hypothesis, we changed the label name of the neutral expression to ‘aneutral’ such that the neutral expression appears first after the labels are one-hot encoded. The

³ One hot encoding is a process whereby categorical variables are converted into a form that can be provided to machine learning algorithms.

violin plots in Fig. 3(d) show the accuracy obtained from neutral and anger expressions before and after the labels were changed. As can be seen from Fig. 3(d), the accuracy obtained before and after the label change was nearly the same or relatively similar for both anger and neutral expression. In addition, no significant difference was found in the achieved accuracies before and after the label was changed. Thus, label order was not important.

Although the achieved accuracy by the computer systems is relatively low when the fully covering masks were used, the above-mentioned observation suggests that the prediction of the computer systems was not fully random. One possible reason why the images of the fully covering masks were mainly classified as anger could be because the produced (extracted) features⁴ by the computer systems when the fully covering masks were used have high similarities with the anger expression images.

Depending on the type of covering, the classification accuracy in humans changes. There is no general order of increase or decrease in the achieved accuracy. However, based on the results obtained in this study, the classification accuracy of the computer systems correlates with the occlusion. The larger the occlusion, the higher the decrease and vice-versa. For instance, in the case of VGG19, the highest accuracy was achieved with nothing added to the image. The achieved accuracy decreases (by up to 18%) when sunglasses (covers 10%–15% of the face) were used. And although a further decrease (up to 65%) has been observed when partial masks (covers 30%–35% of the face) were used, the largest decrease (up to 74.48%) was experienced when the fully covering (covers 60%–65% of the face) face masks were used. It is important to bear in mind that this observation is only valid for the type of coverings used in this research as it is unknown whether the decrease is based on the percentage of the face that is covered or based on important features that are covered. This is an important question for future research.

We also found that people's exposure to people wearing masks (see Section 3.1) does not affect their performance in categorizing emotion. Although more exposure to people wearing mask was related to increased accuracy across all facial expressions, we found no evidence for people's exposure to masked individuals helping them to be specifically better in categorizing masked emotions (see Table 2e). Taken together, these results suggest that people perform worst in categorizing masked emotion images independent of their exposure to masked individuals. In other words, being around more masked people does not increase the ability to categorize partially covered emotional facial expressions.

We trained the computer systems with the last-half frames of images from the CK+ dataset. This means that the computer systems have seen everyone in the dataset (even though testing was performed only on unseen images) whereas human participants had not seen any of those images before. On the other hand, human beings have seen a number of different expressions of people from the time they were born that the computer systems have not seen. Nevertheless, this is on par with recent studies conducted to perform a similar type of comparison as it seems not feasible to train the computer systems with all images the human participants have seen in the past or the other way around.

Prior work [25,26] conducted to tackle the problem of imbalanced training at the data-level⁵ suggests that it is crucial to balance the distribution of data within the train set in order to have a reduced misclassification cost, as well as improve

classification accuracy. However, the findings based on the misclassification analysis suggest that it might be important to feed imbalanced data to classifiers. For instance, in situations like this where using balanced data might lead to differences in misclassifications, feeding the classifier with data that is oversampled based on a certain expression with less cost of incorrect identification (see Section 3.2.1) is more important than using data with an equal distribution from each expression as it results in a high cost of misclassification.

Not only did oversampling reduce the misclassification cost of the covered images when a specified class is targeted, at the same time, it also increased the classification accuracy when classifying the fully covering mask images (in the case of VGG19). The accuracy achieved by the oversampling method (ratio 5:2) was greater than the accuracy achieved when no oversampling was used. Nevertheless, oversampling the data might also come at a cost of decreasing the classification accuracy, e.g. the accuracy obtained from classifying the partial mask and the sunglasses images were reduced after the neutral expression was oversampled for all the three (VGG19, ResNet50, and InceptionV3) algorithms.

This gives rise to the question of which performance measure to prefer? Is the accuracy or the misclassification cost more relevant, taking into account the protective factor, mediating, and moderating effects [53,54] of these measures? While there is no general answer to this question, both the accuracy and the misclassification cost need to be considered. For instance, misclassifying anger or sadness as happiness could have extreme consequences in future decision making for artificial intelligence systems like robots. At the same time, an organization will more likely introduce robots that can accurately classify people's emotions into shared workspaces compared to less accurate ones.

It is also worth stating that the finding that oversampling reduced the cost of misclassification cannot be generalized to all the different types of face coverings. For instance, the misclassification cost with no oversampling for faces with sunglasses was lower (18%) compared to when oversampling (ratio 3:2) was used (23%).

In general, given a set of image features, the computer systems are more likely to predict the emotion category of an image as the same, while humans seem not to show that tendency. For instance, in Section 3.1, without coverings on the face, the computer systems (from all 30 runs) categorized six out of seven emotion categories (except neutral) correctly. However, there is a high variation in the accuracy of humans across emotion categories. An image that is seen to represent a particular emotion by one person might be seen to have a different emotion by another person. This is due to the different perspectives that we all have, and based on the way our brain processes information and mediates actions [55].

Several reports have shown that machine learning classifiers are mainly used nowadays to produce emotion categorization devices [43,56–58]. However, the findings of this research provide insights that these classifiers do not perform well in categorizing emotion when the face is partially covered. Therefore, these types of classifiers should be implemented carefully in contexts where there are coverings on people's faces, especially with the recent challenge of the coronavirus. A possible way to improve the performance of these models is to train them with facial images with coverings.

It is also important to note that the overall findings of what is referred to as computer systems in this research may be somewhat limited to end-to-end deep learning models as different machine learning techniques might lead to achieving different results depending on the feature extraction technique, as well as the algorithm used to categorize the emotion.

Across all images, humans spent approximately 0.14 min to predict the category of an image and 19.77 min for the complete

⁴ Feature extraction is the process by which an initial set of raw data (in this case image pixels) is reduced to subset of features for processing.

⁵ The data-level way of handling imbalanced data tries to balance the training set by increasing the number of samples from smaller classes through the addition of new samples and reducing larger classes.

survey, while it took the computer systems 0.00012 min (less than 1 s) to classify each image and approximately 0.017 min (1 s) to classify all 140 images. Nevertheless, the computer systems took approximately 80 min to train on a 24 GB Graphical Processing Unit (GPU) machine with Quadro RTX 6000 CUDA version 11.2. Thus, computer systems can predict the expression of the images faster than humans provided they are already trained. Humans train throughout their lifetime, although they are already 'experts' at recognizing emotions at an early age [59,60].

5. Conclusion

This study compared the performance of humans and computer systems in categorizing emotion from faces with different coverings. The results obtained from the study showed that while the computer systems (98.48%) outperformed humans (82.78%) with no covering on the face, the performance of the computer systems showed a larger decrease (up to 74.48% decrease) in comparison to humans (with 25.29% decrease) when the face is partially covered. Further, as these computer systems make different mistakes in comparison to humans, this work introduced a method to reduce the cost of misclassification for computer systems in order to avoid having extreme consequences in future decision making. This highlights how partial face coverings affect the performance of humans and computer systems in classifying emotions. The findings contribute to our understanding of how computer systems make mistakes, as well as how the introduced approach reduces the misclassification cost for computer systems. Together, these findings suggest that greater efforts are needed to develop artificial emotion classification systems that can adapt to more complex situations, e.g. taking into account contextual features. Thus, care is needed prior to using results from computer systems for emotion recognition in soft computing experiments.

The study also has some potential limitations. For instance, the overall findings of what is referred to as computer systems in this research may be somewhat limited to end-to-end deep learning models as different machine learning techniques might lead to achieving different results depending on the feature extraction technique, as well as the algorithm used to categorize the emotion. In addition, the study was performed on a single stimulus set, therefore, the question of whether the findings would generalize to other stimulus sets remains open. Future studies should address this issue by including other stimulus sets, as well as other artificial emotion classification methods.

CRediT authorship contribution statement

Harisu Abdullahi Shehu: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization, Project administration. **Will N. Browne:** Supervision, Writing – review & editing. **Hedwig Eisenbarth:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Authors would like to declare a competing interest with Bing Xue and Mengjie Zhang as they are colleagues whom we have a close working relationship with.

Data availability

The CK+ dataset used in this research is publicly available and can be downloaded from <http://www.jeffcohn.net/Resources/> The humans' data can be downloaded from either OSF (<https://osf.io/mgx9p/>).

Acknowledgments

Authors would like to acknowledge the participants for putting in the effort to complete the survey.

Funding

This research was not supported by any organization.

Ethics approval

Ethics approval has been obtained from the Human Ethics Committee (HEC) of Victoria University of Wellington (Application no: 28949).

Consent to participate

All participants gave informed consent for participation in the study.

Consent for publication

All authors have approved the manuscript and agree with submission to Applied Soft Computing journal

Code availability

Code implemented in this study can be downloaded from OSF (<https://osf.io/mgx9p/>) or Code Ocean (<https://doi.org/10.24433/CO.2976335.v1>)

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.asoc.2022.109701>.

References

- [1] C.-C. Carbon, Wearing face masks strongly confuses counterparts in reading emotions, *Front. Psychol.* 11 (2020) 2526.
- [2] D. Roberson, M. Kikutani, P. Döge, L. Whitaker, A. Majid, Shades of emotion: What the addition of sunglasses or masks to faces reveals about the development of facial expression processing, *Cognition* 125 (2) (2012) 195–206.
- [3] L.F. Barrett, How emotions are made: The secret life of the brain, 2017.
- [4] P. Lewinski, T.M. den Uyl, C. Butler, Automated facial coding: Validation of basic emotions and FACS AUs in FaceReader, *J. Neurosci. Psychol. Econ.* 7 (4) (2014) 227.
- [5] E. Noyes, J.P. Davis, N. Petrov, K.L. Gray, K.L. Ritchie, The effect of face masks and sunglasses on identity and expression recognition with super-recognizers and typical observers, *R. Soc. Open Sci.* 8 (3) (2021) 201169.
- [6] M. Wegrzyn, M. Vogt, B. Kireclioglu, J. Schneider, J. Kissler, Mapping the emotional face. How individual face parts contribute to successful emotion recognition, *PLoS One* 12 (5) (2017) e0177239.
- [7] M. Yuki, W.W. Maddux, T. Masuda, Are the windows to the soul the same in the East and West? Cultural differences in using the eyes and mouth as cues to recognize emotions in Japan and the United States, *J. Exp. Soc. Psychol.* 43 (2) (2007) 303–311.
- [8] E.J. Miller, E.G. Krumhuber, A. Dawel, Observers perceive the Duchenne marker as signaling only intensity for sad expressions, not genuine emotion, *Emotion* (2020).
- [9] L.A. Camras, K. Allison, Children's understanding of emotional facial expressions and verbal labels, *J. Nonverbal Behav.* 9 (2) (1985) 84–94.
- [10] M. Guarnera, P. Magnano, M. Pellerone, M.I. Cascio, V. Squatrito, S.L. Buccieri, Facial expressions and the ability to recognize emotions from the eyes or mouth: A comparison among old adults, young adults, and children, *J. Genet. Psychol.* 179 (5) (2018) 297–310.
- [11] J. Wacker, M. Heldmann, G. Stemmler, Separating emotion and motivational direction in fear and anger: Effects on frontal asymmetry, *Emotion* 3 (2) (2003) 167.

- [12] M.G. Calvo, D. Lundqvist, Facial expressions of emotion (KDEF): Identification under different display-duration conditions, *Behav. Res. Methods* 40 (1) (2008) 109–115.
- [13] K. Hugenberg, Social categorization and the perception of facial affect: target race moderates the response latency advantage for happy faces, *Emotion* 5 (3) (2005) 267.
- [14] J. Van den Stock, R. Righart, B. De Gelder, Body expressions influence recognition of emotions in the face and voice, *Emotion* 7 (3) (2007) 487.
- [15] M. Marini, A. Ansani, F. Paglieri, F. Caruana, M. Viola, The impact of facemasks on emotion recognition, trust attribution and re-identification, *Sci. Rep.* 11 (1) (2021) 1–14.
- [16] F. Grundmann, K. Epstude, S. Scheibe, Face masks reduce emotion-recognition accuracy and perceived closeness, *PLoS One* 16 (4) (2021) e0249792.
- [17] H.A. Shehu, W.N. Browne, H. Eisenbarth, An out-of-distribution attack resistance approach to emotion categorization, *IEEE Trans. Pattern Anal. Mach. Intell.* 2 (6) (2021) 564–573.
- [18] R.W. Picard, E. Vyzas, J. Healey, Toward machine emotional intelligence: Analysis of affective physiological state, *IEEE Trans. Pattern Anal. Mach. Intell.* 23 (10) (2001) 1175–1191.
- [19] H.A. Shehu, W.N. Browne, H. Eisenbarth, An adversarial attacks resistance-based approach to emotion recognition from images using facial landmarks, in: 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), IEEE, 2020, pp. 1307–1314.
- [20] H.A. Shehu, A. Siddique, W.N. Browne, H. Eisenbarth, Lateralized approach for robustness against attacks in emotion categorization from images, in: 24th International Conference on Applications of Evolutionary Computation (EvoApplications 2021), Springer, 2021.
- [21] H.A. Shehu, W.N. Browne, H. Eisenbarth, Emotion Categorization from Faces of People with Sunglasses and Facemasks, Springer, 2021, <http://dx.doi.org/10.21203/rs.3.rs-375319/v1>.
- [22] T.J. Coleman, Coronavirus: Call for Clear Face Masks to be 'The Norm', BBC News, 2020, Accessed: 26 May, 2020. [Online]. Available: <https://www.bbc.com/news/world-52764355>.
- [23] N. Japkowicz, et al., Learning from imbalanced data sets: a comparison of various strategies, in: AAAI Workshop on Learning from Imbalanced Data Sets, Vol. 68, AAAI Press, Menlo Park, CA, 2000, pp. 10–15.
- [24] J.W. Grzymala-Busse, Rule induction, in: Data Mining and Knowledge Discovery Handbook, Springer, 2005, pp. 277–294.
- [25] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, *J. Artificial Intelligence Res.* 16 (2002) 321–357.
- [26] E. Ramentol, Y. Caballero, R. Bello, F. Herrera, SMOTE-RS b*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory, *Knowl. Inf. Syst.* 33 (2) (2012) 245–265.
- [27] M. Pazzani, C. Merz, P. Murphy, K. Ali, T. Hume, C. Brunk, Reducing misclassification costs, in: Machine Learning Proceedings 1994, Elsevier, 1994, pp. 217–225.
- [28] M. Kukar, I. Kononenko, C. Grošelj, K. Kralj, J. Fettich, Analysing and improving the diagnosis of ischaemic heart disease with machine learning, *Artif. Intell. Med.* 16 (1) (1999) 25–50.
- [29] M. Ciraco, M. Rogalewski, G. Weiss, Improving classifier utility by altering the misclassification cost ratio, in: Proceedings of the 1st International Workshop on Utility-Based Data Mining, 2005, pp. 46–52.
- [30] J. Berryhill, K.K. Heang, R. Clogher, K. McBride, Hello, World: Artificial intelligence and its use in the public sector, 2019.
- [31] H. Eisenbarth, G.W. Alpers, D. Segré, A. Calogero, A. Angrilli, Categorization and evaluation of emotional faces in psychopathic women, *Psychiatry Res.* 159 (1–2) (2008) 189–195.
- [32] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556).
- [33] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
- [34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016 pp. 2818–2826.
- [35] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, *Adv. Neural Inf. Process.* 25 (2012).
- [36] S. Cheng, G. Zhou, Facial expression recognition method based on improved VGG convolutional neural network, *Int. J. Pattern Recognit. Artif. Intell.* 34 (07) (2020) 2056003.
- [37] K. Mohan, A. Seal, O. Krejcar, A. Yazidi, FER-net: facial expression recognition using deep neural net, *Neural Comput. Appl.* (2021) 1–12.
- [38] Keras API, 2021, [Online]. Available: <https://keras.io/>.
- [39] H.A. Shehu, W.N. Browne, H. Eisenbarth, Emotion categorization from video-frame images using a novel sequential voting technique, in: International Symposium on Visual Computing, Springer, 2020 pp. 618–632.
- [40] P. Ekman, W.V. Friesen, Constants across cultures in the face and emotion, *J. Personal. Soc. Psychol.* 7 (2) (1971) 124, <http://dx.doi.org/10.1109/AFGR.2000.840611>.
- [41] A. Mollahosseini, D. Chan, M.H. Mahoor, Going deeper in facial expression recognition using deep neural networks, in: 2016 IEEE Winter Conference on Applications of Computer Vision, WACV, 2016, pp. 1–10, <http://dx.doi.org/10.1109/WACV.2016.7477450>.
- [42] M. Li, H. Xu, X. Huang, Z. Song, X. Liu, X. Li, Facial expression recognition with identity and emotion joint learning, *IEEE Trans. Affect. Comput.* (2018) 1, <http://dx.doi.org/10.1109/TAFFC.2018.2880201>.
- [43] T. Zhang, W. Zheng, Z. Cui, Y. Zong, Y. Li, Spatial-temporal recurrent neural network for emotion recognition, *IEEE Trans. Cybern.* 49 (3) (2019) 839–847, <http://dx.doi.org/10.1109/TCYB.2017.2788081>.
- [44] H. He, S. Chen, Identification of facial expression using a multiple impression feedback recognition model, *Appl. Soft Comput.* 113 (2021) 107930.
- [45] M.N. Islam, M. Seera, C.K. Loo, A robust incremental clustering-based facial feature tracking, *Appl. Soft Comput.* 53 (2017) 34–44.
- [46] R.J. Cabin, R.J. Mitchell, To Bonferroni or not to Bonferroni: when and how are the questions, *Bull. Ecol. Soc. Am.* 81 (3) (2000) 246–248.
- [47] D. Matsumoto, H.S. Hwang, Evidence for training the ability to read microexpressions of emotion, *Motiv. Emot.* 35 (2) (2011) 181–191.
- [48] E.G. Krumhuber, D. Küster, S. Namba, D. Shah, M.G. Calvo, Emotion recognition from posed and spontaneous dynamic expressions: Human observers versus machine analysis, *Emotion* (2019).
- [49] L.F. Barrett, R. Adolphs, S. Marsella, A.M. Martinez, S.D. Pollak, Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements, *Psychol. Sci. Public Interest* 20 (1) (2019) 1–68.
- [50] E.G. Krumhuber, D. Küster, S. Namba, L. Skora, Human and machine validation of 14 databases of dynamic facial expressions, *Behav. Res. Methods* (2020) 1–16.
- [51] M. Hausmann, B.R. Innes, Y.K. Birch, R.W. Kentridge, Laterality and (in) visibility in emotional face perception: Manipulations in spatial frequency content, *Emotion* 21 (1) (2021) 175.
- [52] L.S. Delicato, R. Mason, Happiness is in the mouth of the beholder and fear in the eyes, *J. Vis.* 15 (1378) (2015).
- [53] R. Rogers, The uncritical acceptance of risk assessment in forensic practice, *Law Hum. Behav.* 24 (5) (2000) 595–605.
- [54] J. Brown, J.P. Singh, Forensic risk assessment: A beginner's guide, *Arch. Forensic Psychol.* 1 (1) (2014) 49–59.
- [55] A. Chakravarti, Perspectives on human variation through the lens of diversity and race, *Cold Spring Harbor Perspect. Biol.* 7 (9) (2015) a023358.
- [56] W.M. Alenazy, A.S. Alqahtani, Gravitational search algorithm based optimized deep learning model with diverse set of features for facial expression recognition, *J. Ambient Intell. Humaniz. Comput.* 12 (2) (2021) 1631–1646.
- [57] D.Y. Choi, B.C. Song, Semi-supervised learning for continuous emotion recognition based on metric learning, *IEEE Access* 8 (2020) 113443–113455, <http://dx.doi.org/10.1109/ACCESS.2020.3003125>.
- [58] W. Hayale, P. Negi, M. Mahoor, Facial expression recognition using deep siamese neural networks with a supervised loss function, in: 2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019), 2019, pp. 1–7, <http://dx.doi.org/10.1109/FG.2019.8756571>.
- [59] S.D. Pollak, D.J. Kistler, Early experience is associated with the development of categorical representations for facial expressions of emotion, *Proc. Natl. Acad. Sci.* 99 (13) (2002) 9072–9076.
- [60] J. Decety, The neurodevelopment of empathy in humans, *Dev. Neurosci.* 32 (4) (2010) 257–267.