European Journal of Orthodontics, 2019, 1–6
doi:10.1093/ejo/cjz007

Original article

Facial attractiveness of cleft patients: a direct comparison between artificial-intelligence-based scoring and conventional rater groups

Raphael Patcas[1], Radu Timofte[2], Anna Volokitin[2], Eirikur Agustsson[2], Theodore Eliades[1], Martina Eichenberger[1] and Michael Marc Bornstein[3]

[1]Clinic of Orthodontics and Pediatric Dentistry, Center of Dental Medicine, University of Zurich, [2]Computer Vision Laboratory, D-ITET, ETH Zurich, Switzerland, and [3]Oral and Maxillofacial Radiology, Applied Oral Sciences, Faculty of Dentistry, The University of Hong Kong, Prince Philip Dental Hospital, Hong Kong SAR, China

Correspondence to: Raphael Patcas, Clinic for Orthodontics and Pediatric Dentistry, Center for Dental Medicine, University of Zurich, Plattenstrasse 11, 8032 Zürich, Switzerland. E-mail: raphael.patcas@zzm.uzh.ch

Summary

Objectives: To evaluate facial attractiveness of treated cleft patients and controls by artificial intelligence (AI) and to compare these results with panel ratings performed by laypeople, orthodontists, and oral surgeons.

Materials and methods: Frontal and profile images of 20 treated left-sided cleft patients (10 males, mean age: 20.5 years) and 10 controls (5 males, mean age: 22.1 years) were evaluated for facial attractiveness with dedicated convolutional neural networks trained on >17 million ratings for attractiveness and compared to the assessments of 15 laypeople, 14 orthodontists, and 10 oral surgeons performed on a visual analogue scale ($n = 2323$ scorings).

Results: AI evaluation of cleft patients (mean score: 4.75 ± 1.27) was comparable to human ratings (laypeople: 4.24 ± 0.81, orthodontists: 4.82 ± 0.94, oral surgeons: 4.74 ± 0.83) and was not statistically different (all Ps ≥ 0.19). Facial attractiveness of controls was rated significantly higher by humans than AI (all Ps ≤ 0.02), which yielded lower scores than in cleft subjects. Variance was considerably large in all human rating groups when considering cases separately, and especially accentuated in the assessment of cleft patients (coefficient of variance—laypeople: 38.73 ± 9.64, orthodontists: 32.56 ± 8.21, oral surgeons: 42.19 ± 9.80).

Conclusions: AI-based results were comparable with the average scores of cleft patients seen in all three rating groups (with especially strong agreement to both professional panels) but overall lower for control cases. The variance observed in panel ratings revealed a large imprecision based on a problematic absence of unity.

Implication: Current panel-based evaluations of facial attractiveness suffer from dispersion-related issues and remain practically unavailable for patients. AI could become a helpful tool to describe facial attractiveness, but the present results indicate that important adjustments are needed on AI models, to improve the interpretation of the impact of cleft features on facial attractiveness.

Introduction

Non-syndromic clefts of the lip and palate are one of the most common congenital defects occurring in approximately 1 in 600 live births worldwide (1). Prolonged interdisciplinary care is indispensable to counteract possible long-lasting negative effects, which clefts would leave on function and appearance. A plethora of therapy approaches have been introduced, including different surgical closure techniques of the clefts, orthodontic treatment, and

orthognathic surgery, all to improve function and facial appearance. Yet regrettably, therapy does frequently not result in an average facial appearance (2), leaving more than often scars from the surgical interventions and an asymmetry around the nose and mouth (3). This reduced facial attractiveness reportedly affects the patients' psychosocial well-being (4), and assessment of facial appearance is therefore considered a crucial factor to measure treatment outcome.

Currently, no valid objective model exists to study the aesthetic treatment outcome in clefts (5), and facial aesthetics is usually established by single raters or panels (6). This tactic is however ill-fated, as it suffers from weaknesses associated with the inconsistency of the subjective raters, which mainly depends on their background (3). To overcome the inhomogeneity seen in subjective ratings, applying a computerized model would seem promising.

In computer science, artificial intelligence (AI) is a term widely used to describe the capability to mimic cognitive functions ordinarily associated with humans, such as learning and problem-solving. For these tasks, convolutional neural networks (CNNs) are frequently designed, which are biologically inspired models and often follow the vision processing in living organism (7). Compared with other image classifications that depend on hand-engineered algorithms, CNNs are subject to very little pre-processing, as they are capable to learn and optimize their output by changing connection weights in relation to the expected result, after each piece of data is processed. CNNs have been applied for various tasks related to facial recognition, spanning from the assessment of facial expressions (7) to the evaluation of apparent age (8), at times clearly outperforming human reference. It is, however, noteworthy that CNNs have not been widely applied to evaluate facial attractiveness. One of the underlying reasons might be that while assessing aspects such as age or gender, objective criteria are used as labels, and datasets are thus relatively easy to collect. Conversely, the attractiveness of a face is essentially subjective and a robust average attractiveness label requires the ratings from a large group of observers. Until very recently (9), the largest face dataset with attractiveness labels contained only 2048 face images (10), and the reported performances were unsatisfying, remaining neither reliable enough nor tuned for medical purposes.

In medicine, the powerful advantages of CNNs in discriminating images have been applied recently—with various success—for instance in melanoma identification (11), sickle cell detection in blood (12), and the automatic classification of ovarian (13) or gastric cancer (14) types. It seems, however, that no CNN has ever been used in the field of dentistry. We, therefore, propose the use of a CNN trained on a very large dataset and fine-tuned for medical assessment, to extract facial features associated with attractiveness. To validate this approach, the present pilot study contains a comparison of CNN to human capabilities.

This investigation aims to provide no more than a proof of concept, verifying the feasibility and practical potential of AI-based ratings in dentistry. To this end, facial attractiveness of treated cleft patients and controls was evaluated by AI and compared with the results of published ratings performed by laypeople, orthodontists, and oral surgeons (3). The hypothesis of this study was that no identifiable differences between the various evaluations exist.

Materials and methods

Materials and subjects

This retrospective analysis was conducted on frontal and left-side profile images retrieved from the archives of the Clinic for Orthodontics and Paediatric Dentistry of the University of Zurich. The test group consisted of 20 files of randomly selected adult patients (10 males, 10 females; mean age: 20.5 years) previously treated inter-disciplinarily for left-sided cleft lip and palate. The treatment protocol included lip repair according to the Millard–Perko protocol, soft palate repair according to Widmaier–Perko protocol, alveolar bone grafting with cancellous bone from the iliac crest, and fixed orthodontic appliances. Orthognathic surgery was conducted in most cases (n = 17), as well as rhinoplasty (n = 14) or minor lip revision (n = 9).

Ten post-retention files of orthodontically treated adults (5 males, 5 females; mean age: 22.1 years) served as control. These were randomly selected cases with pretreatment Angle class I, no major skeletal discrepancies, and treated only for minor dental misalignment. Owing to slightly different recall protocols, the orthodontic records consisted of patients with a discreetly higher average age. Cases were only considered both as tests and as controls when the following criteria could be ascertained: no syndromes, no congenital facial anomalies other than left-sided cleft lip and palate, and no exceptional facial features such as tattoos or piercings.

All patients gave their written informed consent for secondary use of their records including facial photographs. Guidelines in Medical Ethics (15) were strictly obeyed and all data were handled in accordance with State and Federal Law (Human Research Ordinance, Art. 25, §2). Judicial and ethical conformity of this study were confirmed by the governmental ethics committee (BASEC 2016-990; KEK ZH 2012/28).

Methods

The images used consisted of standardized frontal and left-side profile images of each patient, taken 0.5–2 years post-treatment for cleft patients and 3–5 years post-treatment for controls. All photographs (n = 60) were taken with a single-lens reflex camera and a dedicated flash reflector against a monochrome dark blue background, with patients adopting a neutral facial expression in habitual head position. Apart from cropping and brightness or contrast adjustment, no image processing was done. The images were digitally archived and printed for the raters, each image individually and in colour, together with a visual analogue scale (VAS).

The photographs were sent to 20 randomly selected maxillofacial surgeons (20 males; members of the Swiss Society of Oral and Maxillofacial Surgery) with an average age of 56.6 years (range: 43–65 years), 20 randomly selected orthodontists (5 females and 15 males; members of the Swiss Orthodontic Society) with an average age of 51.7 years (range: 28–64 years), and 20 laypersons (5 females and 15 males) with an average age of 52.2 years (range: 34–65 years). Before randomization, care was taken to exclude professional raters affiliated to the local university. Laypeople were selected based on incidental contacts, who were not trained in dentistry, surgery, or aesthetics. The observers were instructed in writing to evaluate the attractiveness of the entire face on the VAS (0: extremely unattractive, 10: extremely attractive) objectively while disregarding external factors such as hairstyle, jewellery, or make-up.

Of the contacted individuals, 15 laypeople (5 females and 10 males), 14 orthodontists (5 females and 9 males), and 10 maxillofacial surgeons (all males) returned their evaluation. Of all possible VAS assessments (n = 2340), 17 were missing and were disregarded for subsequent analyses. Thus, the statistical analyses of the human ratings were based on 2323 independent assessments.

Facial attractiveness was further determined by a computational algorithm consisting of a face detector (16) and CNNs trained to extract facial features associated with attractiveness, to provide a prediction of facial attractiveness (9) (Figure 1).

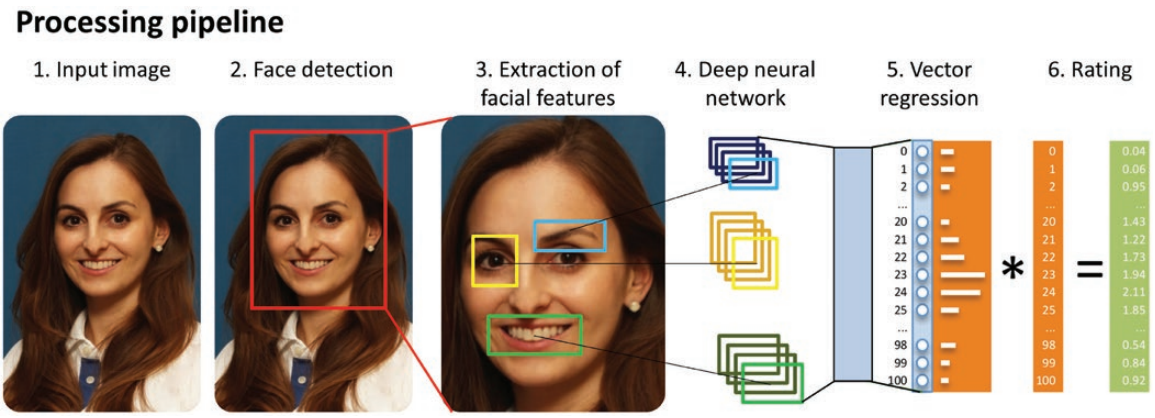

Figure 1. Processing pipeline of the convolutional neural network (CNN) applied. Note that the person depicted was not part of the assessed cases and is shown for illustration purposes only (to indicate that the CNN were not solely trained on standard frontal view, this photograph with a slightly tilted head was selected intentionally).

The Matthias face detector (16) uses a set of learned face templates at different scales that are evaluated in a sliding window manner in the image space and the highest responses are considered face detections. The CNN is an artificial neural network inspired by the visual cortex/vision processing in living organisms. It is defined by a sequence of intermingled layers of neurons, filters, and activation functions comprising parameters taught to map an input face image to the corresponding attractiveness class. The learned CNN model extracts facial features by identifying the important regions in the image, structuring, and combining them.

For this task, all face images were brought to an equal size of 256 × 256 pixels with a centred face and a 40% background margin before being used as input for the CNN models. The CNN models used VGG-16 architecture (17) and were trained on a dataset of a dating site containing >13 000 face images of predominantly Caucasians (age range: 18–37 years, mean age: 25 years) with more than 17 million ratings for attractiveness, performed mostly by users from Switzerland (9). Because images from the dataset were taken in conditions dissimilar to medical assessment, the attractiveness prediction network was further fine-tuned, using frontal neutral face images taken in a controlled setting, from the Chicago Face Dataset (18). The expected attractiveness score was subsequently computed using the following formula:

$$attractiveness = \sum_{i=0}^{3} i \; x \; probability \; of \; class \; i$$

and normalized (scale: 0–10) to enable a direct comparison to VAS scores.

Statistical analysis

Descriptive statistics were calculated and presented for cleft and control patients by different human rating groups and by computational algorithms separately. Data distribution was tested for normality, and, based on the obtained results, parametric *t*-tests or non-parametric Wilcoxon signed-rank tests were performed accordingly, to evaluate whether the difference between human VAS ratings and computational algorithms varied significantly between the three observer groups in cleft patients and control patients separately.

To determine the variation in each human rating group, coefficients of variation (i.e. ratio of standard deviation to the mean) were also estimated as a measure to describe inter-rater variability, by calculating the ratings of each patient separately, for all three rating groups.

The significance level chosen for all statistical tests was $P < 0.05$. All analyses were performed in SPSS (version 24.0; IBM Corp., Armonk, New York, USA).

Results

Descriptive statistics are presented in Table 1 for cleft and control patients of each human rating group and the computational algorithm (AI) used. Normality tests disclosed that the differences between human ratings (all observer groups) and the computational algorithms followed normal distribution (all *P*s > 0.05) except the differences between AI and laypersons and oral surgeons in the cleft subgroup ($P = 0.039$ and 0.038, respectively).

Overall, control cases were judged more favourably (mean: 6.04–7.15) than clefts (mean: 4.24–4.82) by all three human rater groups but not by AI. Detailed analysis of the data (Table 2) revealed that with regard to the evaluation of facial attractiveness for cleft patients no significant differences between human ratings and computational algorithms existed (all *P*s ≥ 0.19), with exceptionally strong agreement between the computational algorithm and both groups of professionals raters (i.e. orthodontists and oral surgeons). Conversely, when evaluating the control group, facial attractiveness was considered significantly higher across all human raters in comparison to the computed results (all *P*s ≤ 0.02).

Coefficient of variation as a precision measure on a per case basis given in Table 1 shows that the variations in human ratings are rather large with mean values ranging from 32.56% to 42.19% in cleft patient images and 23.56% to 29.83% in control patient images. The dispersion was thus particularly large in the assessment of cleft cases, demonstrating the absence of unity in all human rater groups.

Discussion

This research contribution aimed to tackle the existing obstacles in evaluating facial appearance as part of treatment outcome assessment in cleft patients. To this end, facial images of treated cleft patients and controls were fed to a face detector and a previously trained CNN to appraise facial attractiveness. The underlying hypothesis was that no differences would be observable between the obtained results and previously published panel-based ratings (3).

Table 1. Descriptive data for visual analogue scale scores of facial attractiveness of patients treated for clefts and for control patients as assessed by different rating groups and by using computational algorithms (artificial intelligence [AI])

	Layperson (n = 15)	Orthodontist (n = 14)	Oral surgeon (n = 10)	AI[#]
Cleft (n = 40)				
Mean (SD)	4.24 (0.81)	4.82 (0.94)	4.74 (0.83)	4.75 (1.27)
Median (p25, p75)	4.29 (3.73, 4.77)	4.68 (4.29, 5.46)	4.83 (4.21, 5.39)	4.45 (3.81, 5.66)
Minimum	2.20	2.61	2.80	2.74
Maximum	5.90	6.93	6.50	8.44
Imprecision, given as coefficient of variance (COV) per case (%): mean COV (SD COV)	38.73% (9.64%)	32.56% (8.21%)	42.19% (9.80%)	0% (0%)
Control (n = 20)				
Mean (SD)	6.04 (0.53)	7.15 (0.62)	6.49 (0.54)	4.16 (1.04)
Median (p25, p75)	6.05 (5.78, 6.40)	7.20 (6.73, 7.74)	6.35 (6.08, 6.87)	3.96 (3.54, 5.00)
Minimum	4.77	6.14	5.35	2.30
Maximum	7.03	7.96	7.56	5.98
Imprecision, given as: COV per case (%): mean COV (SD COV)	27.18% (6.56%)	23.56% (7.14%)	29.83% (6.29%)	0% (0%)

p25 = 25th percentile; p75 = 75th percentile.
[#]Computational algorithm comprising a face detector and convolutional neural networks.

Table 2. Visual analogue scale scores for facial attractiveness of patients treated for clefts and for control patients: differences between different rating groups to values obtained by trained computational algorithms

	Layperson (n = 15)	Orthodontist (n = 14)	Oral surgeon (n = 10)
Computer			
Difference (cleft)			
Mean (SD)	–0.51 (2.15)	0.07 (2.13)	–0.01 (2.45)
Median (p25, p75)	–0.49 (–2.00, 0.92)	0.19 (–1.32, 1.38)	0.19 (–1.67, 1.86)
Minimum	–8.44	–7.44	–7.94
Maximum	5.79	5.79	5.69
P value	.186	.783	.988
Difference (control)			
Mean (SD)	1.88 (1.94)	2.97 (1.90)	2.32 (2.17)
Median (p25, p75)	1.84 (0.52, 3.27)	3.13 (1.67, 4.27)	2.26 (0.86, 3.89)
Minimum	–3.86	–3.09	–2.52
Maximum	7.20	7.30	7.70
P value	**<0.001**	**<0.001**	**0.002**

p25 = 25th percentile; p75 = 75th percentile.
Bold = Significantly different from zero by mixed model, $P < 0.05$.

The hypothesis had to be partially rejected. Although in cleft patients no differences between the computational algorithms and the different raters were identified, substantial divergences were observed in the assessment of non-cleft patients.

Although the interpretation of this finding is far from obvious, the following explanation might be proposed: in this study, all human raters were asked to assess cleft patients in juxtaposition to non-cleft patients, which probably might have caused a certain bias in focusing on facial characteristics typically attributed to cleft patients. This could have led to an over-interpretation of these facial features. In fact, a previous study compared the evaluation of faces with or without clefts and concluded that cleft patients were looked at with significantly longer fixations around the scared regions and were more negatively rated for appearance compared with the control (19). The authors attributed these findings mostly to residual asymmetry evident in cleft patients. On the other hand, AI might be more robust to countervail this bias. As outlined earlier, the processing pipeline was based on deep neural networks able to extract facial features, which were subsequently weighted and rated using computational algorithms trained on millions of ratings. The applied vector regressions were thus not concentrating on cleft-associated facial features, and it can therefore be assumed that the computational assessment provided a far more impartial and balanced rating. Nevertheless, it is somewhat surprising to realize that AI apparently failed to detect compromising traits in clefts, rating controls, and cleft patients similarly, in contrast to the human rating groups. Understanding certain aspects of the used AI model might help to shed light on this issue: the AI model was not trained to rate specific facial characteristics or particular regions in the image, but rather to undertake a holistic interpretation of the input face image. Within the training process, it learned to weigh the image contents differently, being capable to self-organize and pick up the essential information. The analysis of the activations in the neural network revealed that AI interprets—very much like a heat-map—facial composition rather than particular single traits or features such as hairstyle, background, or jewellery (20). It is, therefore, reasonable to assume this being the cause why AI underestimates the impact of cleft features on facial attractiveness.

Another important finding of this study with regard to repeatability and reproducibility of data is related to the variance seen in each panel. Especially relevant is the observation of high inter-rater

variability, when evaluating the coefficient of variation for each case. The panel-based mean values of coefficient of variation, ranging from 32.56% to 42.19% in cleft patient images, are considerably high and accordingly inherently problematic, as they demonstrate a blatant absence of unity within the rater groups. One of the obvious benefits of a single AI-based score would therefore doubtlessly be the elimination of the evident variability and subjectivity compromising panel-based ratings. Thus, the fact that AI-based results were comparable to the average scores in clefts of all three rating groups (with especially strong agreement to both professional panels) seems to indicate that AI could replace panel ratings while circumventing dispersion-related issues.

Moreover, panel-based scores are usually unavailable for individual case planning. As such, the availability of a dependable AI-scoring could prove to be a welcome diagnostic tool, enabling clinicians to analyse objectively the outcome of surgical procedures and assisting them in discerning favourable aesthetic results.

Analysis of medical images using computers is not entirely of recent vintage, and early attempts of computerized analysis and diagnostics of medical images were made already in the 1960s (21–23). In radiology, data analysed quantitatively by computer as aid to make the diagnostic decisions have been common practice already for some time. Yet in stark contrast to historical advances, the latest introduction of trained CNNs producing an automated computer diagnosis aims not to assist, but rather to replace the radiologist. Apart from the ethical and moral concerns involved, this approach would require a very high-performance level of the computer output, equal to or better than the performance achieved by radiologists.

Regarding AI-based diagnostics in the field of maxillofacial radiology, scientific literature is scarce and embryonic. Primitive neural networks were either applied to compare automatic cephalometric diagnosis for orthodontic treatment to the results obtained from three orthodontic specialists (24) or to evaluate automatic detection of vertical root fractures with intraoral and cone beam computed tomography images using a sample of artificially fractured teeth (25). Although these previous reports attested a high degree of accuracy of AI-based diagnosis, they all rely on rudimentary neural networks operating on task-specific algorithms. This study is the first attempt to introduce *CNNs* based on deep learning and constitutes an important improvement that has produced—in other fields—results comparable to and in some cases superior to human experts (26–28).

Limitations

All three human rater groups differed in size. Initially, 20 individuals were contacted for every group to participate, but the response rate was unequal. Although this in itself did not affect the planned statistical analyses, it limited the amount of data available for analysis. On the basis of the assumption that males have a different aesthetical perception than females, the overrepresentation of males in this study could potentially have affected the results.

Caution should be exercised when comparing the computed results to the different panels assessing facial attractiveness, as comparing panel ratings to computer-generated results cannot validate the latter. Attractiveness is usually defined as the quality to cause interest and desire in the observer, and as such, subjectivity is an inherent part of the definition. As every panel is but a representation of its observer, may they be professionals or laypeople, it would be unsound to validate one panel through the results of another. Likewise, measurements based on AI are a representation of a particular opinion that cannot be validated through comparison. The CNN-based results thus do not replace other panels but rather represent to some degree 'social attractiveness', i.e. the quality to cause 'social' interest and desire. On the basis of millions of ratings retrieved from a dating site, validated and fine-tuned on medical images, the proposed CNN is unquestionably a fitting tool to mirror social opinion of treated patients. And perhaps this is what should be considered most important for patients as treatment outcomes should not be measured by specific panels but by how society views the aesthetic results.

In this investigation, the applied scoring model rated the facial attractiveness of the controls similarly to the facial attractiveness of cleft patients. As already mentioned earlier, this finding is very likely to be a result of the CNN model used and thus warrants further refinement of the computational algorithm applied. The CNN tested was apparently unable to detect facial features that render cleft patients less attractive (in the eyes of human raters). Great effort was made to train the CNN on more than 13 000 images of different individuals of a population containing subjects with clefts, yet the CNN remained clearly unfit to detect and estimate the impact of cleft features to the same degree as humans do. Fine-tuning the algorithm on labelled cleft cases would have surely help to overcome this shortcoming, but such collections are unfortunately unavailable. Because the intention of this pilot study was to provide a proof of concept for a novel computational rating system, this must, at this stage, be partially rejected. Although the potential benefits became evident through direct comparison to over 2000 human ratings, the results clearly underline that any clinical recommendation of the algorithm tested would be premature.

Conclusions

This study introduces a novel method in dental medicine to rate facial attractiveness, by a face detector and a dedicated CNN. Although the introduced method based on AI has the potential to overcome the shortcomings related to variance and inconsistency prevalent in panel assessments, this study made it evident that the presented AI-based scoring is in need of further perfection and refinement to differentiate cleft features of the face that negatively influence the human perception of attractiveness.

Acknowledgement

The authors are grateful to Ms Kar Yan Li, Centralised Research Laboratories, Faculty of Dentistry, The University of Hong Kong, for her valuable assistance regarding the statistical analysis.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Mossey, P.A., Little, J., Munger, R.G., Dixon, M.J. and Shaw, W.C. (2009) Cleft lip and palate. *Lancet (London, England),* 374, 1773–1785.
2. Pruzinsky, T. (1992) Social and psychological effects of major craniofacial deformity. *Cleft Palate-Craniofacial Journal,* 29, 578–84; discussion 570.
3. Eichenberger, M., Staudt, C.B., Pandis, N., Gnoinski, W. and Eliades, T. (2014) Facial attractiveness of patients with unilateral cleft lip and palate and of controls assessed by laypersons and professionals. *European Journal of Orthodontics,* 36, 284–289.
4. Sinko, K., Jagsch, R., Prechtl, V., Watzinger, F., Hollmann, K. and Baumann, A. (2005) Evaluation of esthetic, functional, and quality-of-life outcome in adult cleft lip and palate patients. *Cleft Palate-Craniofacial Journal,* 42, 355–361.

5. Gkantidis, N., Papamanou, D.A., Christou, P. and Topouzelis, N. (2013) Aesthetic outcome of cleft lip and palate treatment. Perceptions of patients, families, and health professionals compared to the general public. *Journal of Cranio-Maxillo-Facial Surgery,* 41, e105–e110.
6. Sharma, V.P., Bella, H., Cadier, M.M., Pigott, R.W., Goodacre, T.E. and Richard, B.M. (2012) Outcomes in facial aesthetics in cleft lip and palate surgery: a systematic review. *Journal of Plastic, Reconstructive and Aesthetic Surgery,* 65, 1233–1245.
7. Matsugu, M., Mori, K., Mitari, Y. and Kaneda, Y. (2003) Subject independent facial expression recognition with robust face detection using a convolutional neural network. *Neural Networks,* 16, 555–559.
8. Rothe, R., Timofte, R. and Van Gool, L. (2015) DEX: Deep EXpectation of Apparent Age from a Single Image. *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pp. 252–257.
9. Rothe, R., Timofte, R. and Van Gool, L. (2016) Some like it hot—visual guidance for preference prediction. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5553–5561.
10. Gray, D., Yu, K., Xu, W. and Gong, Y. (2010) *Predicting Facial Beauty without Landmarks*. Springer: Berlin, Heidelberg.
11. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M. and Thrun, S. (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature,* 542, 115–118.
12. Xu, M., Papageorgiou, D.P., Abidi, S.Z., Dao, M., Zhao, H. and Karniadakis, G.E. (2017) A deep convolutional neural network for classification of red blood cells in sickle cell anemia. *PLoS Computational Biology,* 13, e1005746.
13. Wu, M., Yan, C., Liu, H. and Liu, Q. (2018) Automatic classification of ovarian cancer types from cytological images using deep convolutional neural networks. *Bioscience Reports*, 38, 3.
14. Sharma, H., Zerbe, N., Klempert, I., Hellwich, O. and Hufnagl, P. (2017) Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology. *Computerized Medical Imaging and Graphics,* 61, 2–13.
15. World Medical Association. (2013) World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. *JAMA,* 310, 2191–2194.
16. Mathias, M., Benenson, R., Pedersoli, M. and Van Gool, L. (2014) Face detection without bells and whistles. *European Conference on Computer Vision* 8692, 720–735.
17. Simonyan, K. and Zisserman, A. (2014) Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
18. Ma, D.S., Correll, J. and Wittenbrink, B. (2015) The Chicago face database: a free stimulus set of faces and norming data. *Behavior Research Methods,* 47, 1122–1135.
19. Meyer-Marcotty, P., Gerdes, A.B., Reuther, T., Stellzig-Eisenhauer, A. and Alpers, G.W. (2010) Persons with cleft lip and palate are looked at differently. *Journal of Dental Research,* 89, 400–404.
20. Agustsson, E., Timofte, R., Escalera, S., Baro, X., Guyon, I. and Rothe, R. Apparent and real age estimation in still images with deep residual regressors on APPA-REAL database. In Automatic Face & Gesture Recognition (FG 2017), 12th IEEE International Conference. 2017. IEEE.
21. Lodwick, G.S., Haun, C.L., Smith, W.E., Keller, R.F. and Robertson, E.D. (1963) Computer diagnosis of primary bone tumors. *Radiology*, 80, 273–275.
22. Meyers, P.H., Nice, C.M. Jr, Becker, H.C., Nettleton, W.J. Jr, Sweeney, J.W. and Meckstroth, G.R. (1964) Automated computer analysis of radiographic images. *Radiology,* 83, 1029–1034.
23. Winsberg, F., Elkin, M., Josiah Macy, J., Bordaz, V. and Weymouth, W. (1967) Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis. *Radiology*, 89, 211–215.
24. Mario, M.C., Abe, J.M., Ortega, N.R. and Del Santo, M. Jr. (2010) Paraconsistent artificial neural network as auxiliary in cephalometric diagnosis. *Artificial Organs,* 34, E215–E221.
25. Johari, M., Esmaeili, F., Andalib, A., Garjani, S. and Saberkari, H. (2017) Detection of vertical root fractures in intact and endodontically treated premolar teeth by designing a probabilistic neural network: an ex vivo study. *Dento Maxillo Facial Radiology,* 46, 20160107.
26. Ciresan, D., Meier, U., Schmidhuber, J. (2012) Multi-column deep neural networks for image classification. 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3642–3649.
27. Krizhevsky, A., Sutskever, I., Hinton, G. (2012) ImageNet classification with Deep Convolutional Neural Networks. NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems, 1, pp. 1097–1105.
28. Marblestone, A.H., Wayne, G. and Kording, K.P. (2016) Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience,* 10, 94.