Implementation of artificial intelligence in chronological age estimation from orthopantomographic X-ray images of archaeological skull remains *

• Luka Banjšak (1), Denis Milošević (2), Marko Subašić (2) •

1 – Independent researcher, Zagreb, Croatia

2 – Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia

Address for correspondence:

Denis Milošević
Faculty of Electrical Engineering and Computing
University of Zagreb, Croatia
E- mail: denis.milosevic@fer.hr

Bull Int Assoc Paleodont. 2020;14(2):122-129.

Abstract

One of the primary steps in forensic dental analysis is age estimation. Alongside sex estimation, this is offers basic categorization of subjects. Whether it is used in person-identification or archaeological analysis and research, a forensic dentist will observe these parameters when starting his work. Orthopantomographic x-ray images offer a lot of data and basically represent the golden standard for identification in forensic stomatology. Deep convolutional neural networks are establishing their presence in numerous fields of medicine and therefore we have explored the possibility of their implementation in age estimation in forensic dentistry. We developed a deep convolutional neural network, based on a dataset of 4035 orthopantomographic images, captured by and kindly provided by University of Zagreb's, School of Dental medicine. A quick, automated and accurate model was formed that opens a new door in the field of forensic dentistry. The developed convolutional neural network was used to estimate the age of 89 archaeological skull remains. The skulls were scanned with an orthopantomography x-ray machine and the received images were used as a testing dataset. The results offered a noteworthy 73% accuracy of placing the images in correct age groups.

Keywords: forensic odontology; convolutional neural network; orthopantomography

** Bulletin of the International Association for Paleodontology is a journal powered by enthusiasm of individuals. We do not charge readers, we do not charge authors for publications, and there are no fees of any kind. We support the idea of free science for everyone. Support the journal by submitting your papers. Authors are responsible for language correctness and content.*

Introduction

Specimen age determination is one of the primary forensic premises. (1) Analysis of the jaw and teeth represent the golden standard for age calculation due to the fact it is non-invasive and offers permanence and repeatability. (2) Due to the fact that dental structures and the head are one of the most durable and best protected parts of the human body, they resist decomposition and are usually the last to deteriorate postmortem. (3) In addition, this method is applicable to both living and dead subjects and merely requires a recent radiographic image of both jaws and teeth (orthopantomographic image), (4).

Manual age determination of adults via one orthopantomographic image is somewhat difficult and rather time consuming. (5) It requires precise measurements, data comparison and understanding of some advanced dental postulates. An average dentist is not trained to do this, so there is a need for specialized personnel, with extensive training. (6)

Comparable to all other computer implementations, with the help of computer hardware we can drastically speed up the process. Virtually within miliseconds a convolutional neural network can determine the approximate age of a person if provided with an orthopantomographic image. With further development of this pilot method there is potential to precisely identify the person's age using just a radiographic image.

The possibility for implementation of artificial intelligence in all fields of medicine is on a daily rise. (7) Artificial intelligence provides a way to automatically discover features and generalize based on the data it has already encountered, achieving noteworthy performance on new and unseen data within the same domain. (8) The effectiveness of neural networks is as high as the time and frequency of their usage. This includes feeding the convolutional neural networks (CNNs) more similar data to gain experience on- also known as training; very much like conventional human intelligence. Naturally the more it is trained, the more accurate it will be.

Over the course of the last decade deep learning went from theoretical discussions to ground-breaking achievements throughout science and industry. This change was felt particularly in the field of computer vision, defined by the "ImageNet moment"(9), where the work by Krizhevsky et al. (10) achieved a 41% better performance on the ImageNet Large Scale Visual Recognition Challenge in 2012 than the next best competitor. While discussions about the feasibility of deep convolutional neural networks (DCNNs) were ongoing for decades, and as the field itself went through a phase now called "AI winter", the lack of large amounts of data and the lack of processing power as well as large-scale storage kept those discussions theoretical. The hardware breakthrough came in the form of dedicated graphics processors (GPU cards), due to their specialized hardware for highly parallelized operations, like matrix multiplication which forms the basis of modern deep learning.

Deep neural networks are general function approximators which are inspired by biological neural processes. They are, in a manner of speaking, black-box models that can approximate any function, but their structure does not give insight into the structure of the function being approximated-meaning, it's difficult to comprehend. Interpretability of deep neural networks is an active field of research, yielding methods like GradCAM (11) for determining significant regions of input images, and LIME (12) that uses local surrogates to explain individual predictions.

Deep learning and other AI methods are being applied to more and more problems in the medical domain every day (13) (14). Some examples include hemorrhage detection in fundus images (15), sex assessment from panoramic dental x-ray images in adults (16) and diagnosis of diabetic retinopathy (17).

Inside the dental field neural networks have been reported since 2017, starting with Miki et al. who combined Cone beam computed tomography (CBCT) and convolutional neural networks (CNNs).(18) Furthermore, CNNs were used in the fields of endodontics, periodontology and cariology.(19) (20) (21) A very notable study by Tuzoff et al. describes the utilization of CNNs in tooth detection with orthopantomography.(22) Schwendicke et al. have explored the possibilities of CNN usage in dental diagnostics. Matsuda, S.et al. have proven that CNNs are capable of identifying 2 separate x-ray images of the same person with 100% accuracy in some cases. (23) That paper, however, did not have a test group, but more on that later.

One significant recent example is the just announced AlphaFold 2 (24), a solution to the protein folding problem (25) (26), with an improvement so significant that many describe it

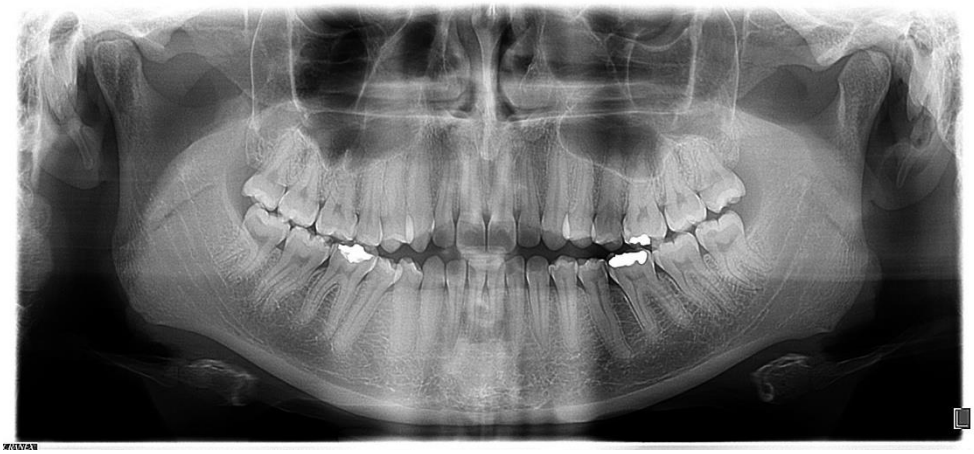

Figure 1 A sample of an orthopantomography x-ray image used to train, validate and test the models.

as "the ImageNet moment" for biology. AI methods are spreading to all research fields, be they in the form of improvement in accuracy and usability or revolutionization of a field like AlphaFold 2, as they allow researchers to gain insight for gigantic datasets with an even larger set of features that would be impossible to analyze manually.

Materials and Methods

In this study a dataset of 4035 panoramic dental x-ray images of living subjects was used, with the ages ranging from 19 to 85 years, with an average age of 38 years. (Figure 1) The female to male ratio is 58.7% to 41.3%. The samples are collected from multiple locations in Croatia and belong to the collection of the Department of Dental Anthropology School of Dental Medicine University of Zagreb. The use of this collection for research purposes has been approved by the School of Dental Medicine University of Zagreb Ethics committee. The images were taken by a wide range of orthopantomography x-ray machines. The data was split into train/test/validation sets, with the ratio of their size being 80% to 10% to 10%, respectively. We used VGG16 (27) architecture pre trained on ImageNet as the feature extractor, which is extended by one 1x1 convolutional layer with 40 channels, which is followed by a fully connected layer of 128 units, ending the deep neural network in a single fully connected unit that estimates the age. All activation used are ReLU (28). VGG16 consists of 5 blocks, with the first two having two convolutional layers followed by max pooling, and the latter three having three convolutional layers followed by max pooling. (Figure 2) The data specimen that we've used the CNN on consists of 89 orthopantomographic images of archaeological skull remains, deceased anytime between the 8th and 11th century and (Figure 3) originating from Croatia. The approximative age of archaeological skull remains was determined from archaeological records and a cross check verification done according to current forensic standards. (5) Those values were sorted in 4 age groups: 0-15 years, 16-30 years, 31-60 years and 61+years and then compared with the ones provided by the CNN.

To determine the best architecture for this problem, experiments have been run on all state-of-the-art architectures, which includes DenseNet201 (29), ResNet50 (30), VGG16, VGG19 (27) and Xception (31). The base network was unchanged, but the end layers were replaced by a 1x1 convolution of size 40 and a fully connected layer with 128 units, as stated before. (Figure 4.) After those preliminary experiments where hyperparameters were evaluated, a fine-tuning step was performed on the best performing model to further improve performance.

It is important to point out that transfer learning was used. Transfer learning is a method used in deep learning where instead of using randomly initialized layer weights, pre-trained layer weights are used. Those weights were trained on the gigantic ImageNet dataset. As mentioned before, overfitting is a big problem when dealing with big models. Tuning 50 million parameters based on 4035 images leads to overfitting, which was noticeable due to the good evaluation performance on the train set, but increasingly worse results on the validation set. The main motivation for using layer weights obtained from

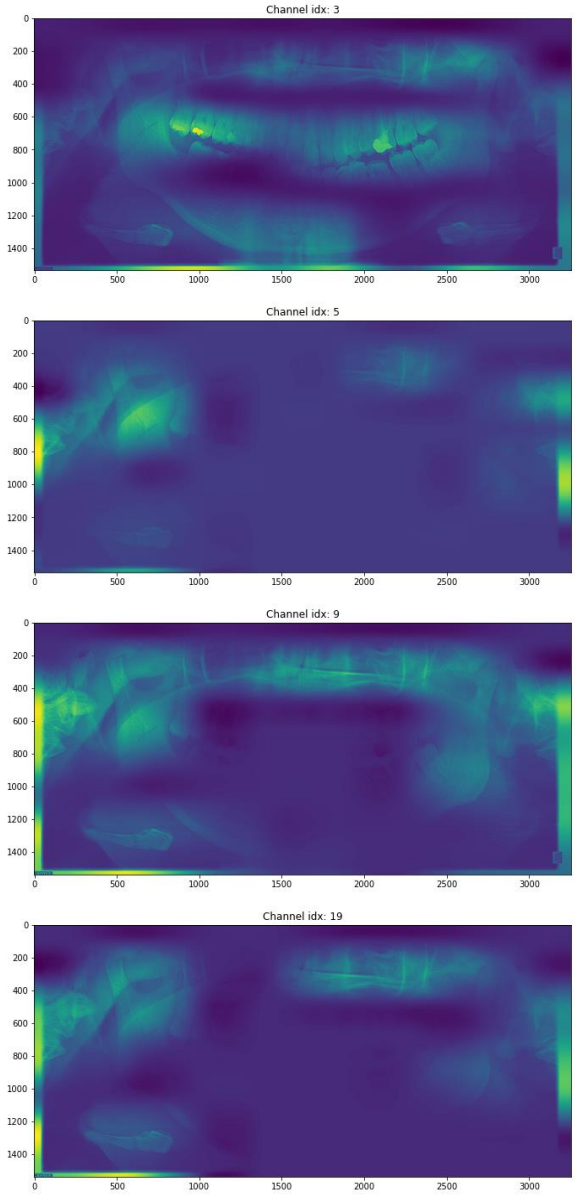

Figure 2 A visualization of activations of the final convolutional layer. The final convolutional layer consists of 40 channels (4 shown on figure) which represent features that the network discovered as useful for age estimation. While no valid interpretation of the model's behavior can be constructed solely from these internal representations, those images give an insight into which components of the image the model regards as a region of interest.

a general-purpose dataset instead of a medical dataset is that: 1) there is no open dataset medical dataset, 2) while higher level features are not useful for medical images (for example, regions of the neural network might "react" to fur, eyes, basketballs or other "less abstract" patterns), the lower level features are usually useful, with some even corresponding to activations measured in the human visual system (circles, edges, checkered patterns and such).

Another important point is the difference in data on which the model was trained on, and data on which the model was evaluated on. The model was trained on panoramic dental x-ray images of living subjects of varying ages, living in contemporary times, with their skulls intact. However, the evaluation was done on deceased subjects ranging from 8th to the 11th century, some with damage to the skeletal and dental structure often incompatible with life. In addition, living subjects are usually aligned in the orthopantomography x-ray machine by biting down, which proves to be difficult to reproduce for deceased subjects. It is reasonable to assume that better performance could be achieved by acquiring a larger dataset, either by introducing deceased subjects to the dataset or by acquiring a much larger dataset of living subjects with the intention of finding better features to estimate the age from. In addition, evolutional differences and dental medicine practice protocol differences are also aggravating circumstances.

Results

We ran the experiment 3 times, each time slightly changing the hyperparameters in order to explore in which case the results would be the most accurate. These runs will be referred to as models. Model 1 was composed as a preliminary high-capacity model. Model 2 is the precisely described one in the “materials” section of this paper, with a mild model stability intervention. Finally, model 3 was architecturally identical to model 2, but with extensive model stability optimization.

In the first analysis model 1 was used. Out of 89 images, 47 of them were accurately placed in respective age groups, while 42 weren’t. This offers an accuracy of 53%. In the second analysis model 2 was used. Out of 89 images, 37 of them were accurately placed in respective age groups, while 52 weren’t. This being the worst run offering an accuracy of just 42%.

In the third and final analysis model 3 was used. We can see that out of 89 images 65 of them were accurately placed in respective age groups, while 24 weren’t. This offers a highly suitable accuracy of 73%. (Figure 5)

The results were delivered in a matter of 1.47 seconds, on conventional, consumer-grade hardware. This is an outstanding result that opens the door to utilization of this software on

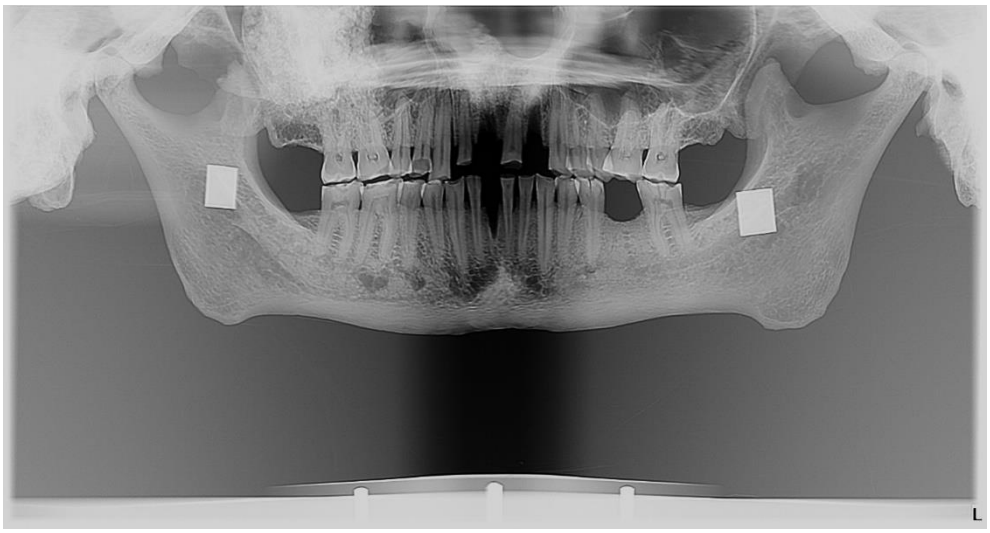

Figure 3 A sample of an orthopantomography x-ray image of an archeological skull remain.

any personal computer, smartphone or tablet; provided it is optimized for the operating system beforehand.

Discussion

The results yield the way to further exploration and utilization of AI in the field of forensic dental medicine. New methodologies are usually met with a healthy dose of skepticism, as it should be. It is hard to incorporate a method that can't be fully explained, especially in the medical field where the chain of conclusions for a diagnosis is imperative. While research is being done in the field of interpretability, current models are evaluated with rigorous examination of their results. The dataset is usually divided into three different parts - the train set, the validation set and the test set. The train set, as the name implies, is used to train the network (to tune the weights of all layers). The validation set is not used for training, but it is used during training. The model is at each step evaluated on the validation set, giving insight into the performance of the model on unseen data. That's an important aspect of deep model training, as overfitting can occur. Overfitting is the phenomena where the model memorizes all input-output pairs and performs well on those, while being unable to produce usable results on unseen data. The results on the validation set are used for research decisions, like changing hyperparameters (for example the architecture of the network), the preprocessing of the data and other things. The validation set is, in a matter of speaking, a reference point that offers guidance in the right direction. We repeat this process until the results achieve a satisfying value. Only then do we evaluate our model on the test data set. The reported performance is only the performance on the test dataset, which in no way contributed to the training process or hyperparameter choices. Only then can it be claimed that the model generalized and that the model is viable for use.

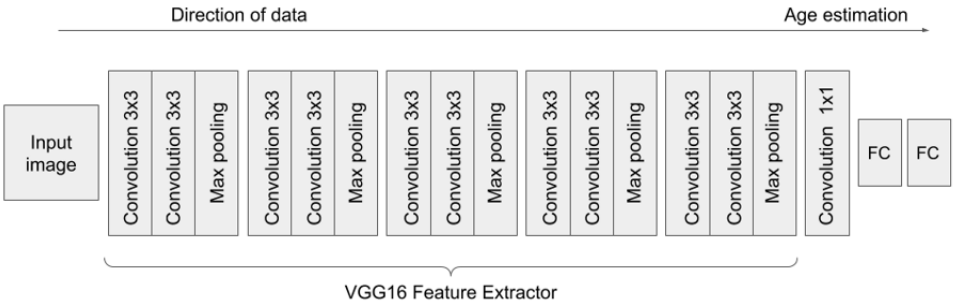

Figure 4 A schematic image of the developed deep convolutional neural network.

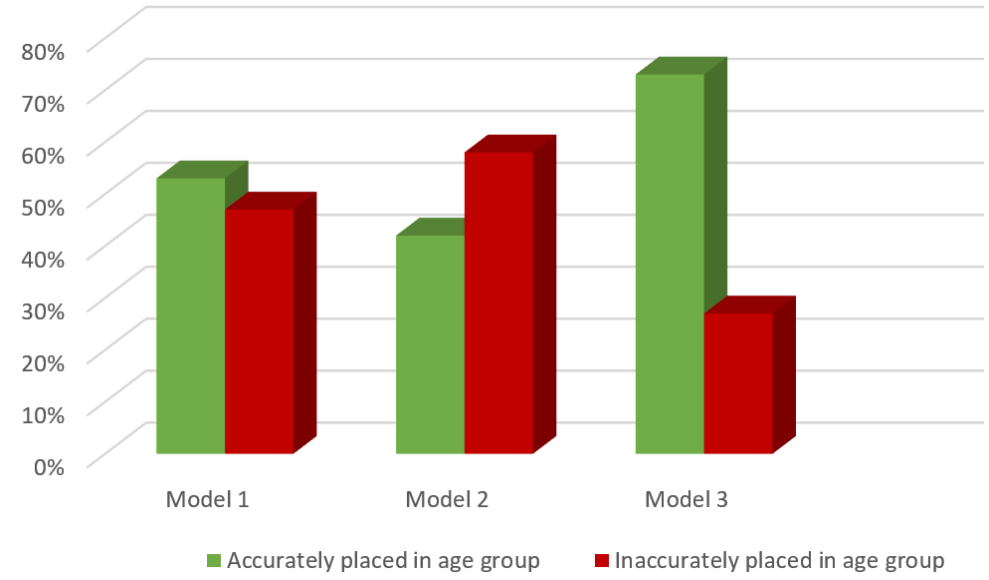

Figure 5 Visual representation of model accuracy.

Conclusion

AI has great potential in the medical field. The more standardized the observed specimen is the more accurate the results will be. Even though the convolutional neural networks can, in fact, belay the imperfections to an extent, in some cases they can cause an imprecise result. For example, in the real-world forensic dentists often deal with poor resolution images, radiological artifacts, blurry images etc. It has become increasingly popular to scan physical images by merely setting them on a surface and taking photographs with smartphones. This introduces color shifting, various white balance settings and image transcoding which additionally decreases the image quality. If such an image is presented to the CNNs it produces disturbances in the input data. The previously mentioned term "Model stability" describes how well a model performs when faced with minor disturbances in the input data, with models being more stable if they perform well despite those perturbations. Model stability can be deliberately targeted during the training by introducing noise akin to the noise expected in the real data (called data augmentation), by acquiring a larger dataset, by extending the capacity of the model and so on. Despite all that, no model is perfect.

Extremely precise age determination is difficult even for an experienced forensic dentist, let alone a convolutional neural network. It is certain that with the further development and training of this neural net would offer better and more precise results. For that an extensive network of radiographic images is needed, alongside the proprietary data: age, sex, morphological differences explanations etc. Naturally this is a tremendous amount of work that needs physical human manual input and "correction", as well as ethical and data protection approvals. That being said, with every experiment the neural net receives additional data, variables and value.

In a group of 89 images in total, this neural net has succeeded to determine the age of all images with a precision rate of 73%. This already puts the experiment in the area of usable methods that may be considered in analysis of panoramic dental x-rays.

It is important to note what neural networks actually do. They are not only finding relations between features and biomedical parameters. They approach the data without prior knowledge, finding the features and parameters themselves from which conclusions are then drawn. In this paper, we have given an overview of what are neural networks, how they work and analyze images and how they can evolve over time. In addition, we observed the pros and cons of machine learning in a realistic situation, often needed in real world forensic stomatology.

References

1. Sweet D. Why a dentist for identification? Dental Clinics of North America. 2001 Apr 1;45(2):237-51.
2. Higgins D, Austin JJ. Teeth as a source of DNA for forensic identification of human remains: a review. Science & Justice. 2013 Dec 1;53(4):433-41.

3. Pretty IA. (2007). Forensic Dentistry: 1. Identification of Human Remains. Dent Update 2007; 34:621-634
4. Alsaffar H, Elshehawi W, Roberts G, Lucas V, McDonald F, Camilleri S (2017) Dental age estimation of children and adolescents: validation of the Maltese Reference Data Set. J Forensic Leg Med 45:29–31
5. Smith T, Brownlees L (2011) Age assessment practices: a literature review and annotated bibliography. United Nations Children's Fund (UNICEF), New York
6. Forrest, A. S. Collection and recording of radiological information for forensic purposes. Aust. Dent. J.57, 24–32 (2012)
7. IEEE Board of directors. Artificial Intelligence: IEEE position statement. IEEE Advancing Technology for Humanity (2019).
8. Russell, S. and Norvig, P., 2002. Artificial intelligence: a modern approach.
9. Deng, Jia & Dong, Wei & Socher, Richard & Li, Li-Jia & Li, Kai & Li, Fei Fei. (2009). ImageNet: a Large-Scale Hierarchical Image Database. IEEE Conference on Computer Vision and Pattern Recognition. 248-255. 10.1109/CVPR.2009.5206848.
10. Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In Pereira, F.; Burges, C.; Bottou, L.; and Weinberger, K., eds., Advances in Neural Information Processing Systems 25. Curran Associates, Inc. 1097–1105.
11. R.R.Selvaraju, M.Cogswell, A.Das, R.Vedantam, D.Parikh, and D.Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), Oct 2017.
12. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?" explaining the predictions of any classifier," in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135–1144, 2016.
13. W. Shen, M. Zhou, F. Yang, C. Yang, and J. Tian, "Multi-scale Convolutional Neural Networks for Lung Nodule Classification," in Information Processing in Medical Imaging (S. Ourselin, D. C. Alexander, C.-F. Westin, and M. J. Cardoso, eds.), Lecture Notes in Computer Science, pp. 588–599, Springer International Publishing, 2015.
14. Litjens G, Kooi T, Bejnordi BE, Setio AA, Ciompi F, Ghafoorian M, Van Der Laak JA, Van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Medical image analysis. 2017 Dec 1;42:60-88.
15. M. J. J. P. v. Grinsven, B. v. Ginneken, C. B. Hoyng, T. Theelen, and C. I. Sánchez, "Fast Convolutional Neural Network Training Using Selective Data Sampling: Application to Hemorrhage Detection in Color Fundus Images," IEEE Transactions on Medical Imaging, vol. 35, pp. 1273– 1284, May 2016.
16. J.Xin, Y.Zhang, Y.Tang, and Y.Yang, "Brain Differences Between Men and Women: Evidence From Deep Learning," Frontiers in Neuroscience, vol. 13, 2019.
17. Pratt H, Coenen F, Broadbent DM, Harding SP, Zheng Y. Convolutional neural networks for diabetic retinopathy. Procedia Computer Science. 2016 Jan 1;90:200-5.
18. Miki, Y. et al. Classification of teeth in conebeam CT using deep convolutional neural network. Comput. Biol. Med.80, 24–29 (2017).
19. Lee, J. H., Kim, D. H., Jeong, S. N. & Choi, S. H. Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. J. Dent.77, 106–111 (2018).
20. Krois, J. et al. Deep learning for the radiographic detection of periodontal bone loss. Sci. Rep.9, 8495 (2019).
21. Ekert, T. et al. Deep learning for the radiographic detection of apical lesions. J. Endod.45, 917-922.e5 (2019).
22. Tuzoff, D. V. et al. Tooth detection and numbering in panoramic radiographs using convolutional neural networks. Dentomaxillofac. Radiol.48, 20180051 (2019).
23. Matsuda, S., Miyamoto, T., Yoshimura, H. et al. Personal identification with orthopantomography using simple convolutional neural networks: a preliminary study. Sci Rep 10, 13559 (2020)
24. Callaway E. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. Nature. 2020 Nov 30.(cited Nov.2020.) doi: https://doi.org/10.1038/d41586-020-03348-4
25. H.S.Chanand K.A.Dill, "Theproteinfoldingproblem," Physicstoday, vol. 46, no. 2, pp. 24–32, 1993.
26. K. A. Dill and J. L. MacCallum, "The protein-folding problem, 50 years on," science, vol. 338, no. 6110, pp. 1042–1046, 2012.
27. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in International Conference on Learning Representations, 2015.
28. V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in ICML, 2010.
29. G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proceedings of the IEEE Confer- ence on Computer Vision and Pattern Recognition (CVPR), July 2017.

30. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.
31. F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (Honolulu, HI), pp. 1800–1807, IEEE, July 2017.