Dental Anomaly Detection Using Intraoral Photos Via Deep Learning

Ronilo Ragodos
University of Iowa

Tong Wang
University of Iowa

Carmencita Padilla
National Institutes of Health, University of the Philippines Manila

Jacqueline Hecht
University of Texas Health Science Center at Houston

Fernando Poletta
ECLAMC at Center for Medical Education and Clinical Research, CEMIC-CONICET

Iêda Orioli
Federal University of Rio de Janeiro

Carmen Buxó
University of Puerto Rico

Azeez Butali
University of Iowa

Consuelo Valencia-Ramirez
Clinica Noel

Claudia Restrepo Muñeton
Clinica Noel

George Wehby
University of Iowa

Seth Weinberg
University of Pittsburgh

Mary Marazita
University of Pittsburgh

Lina Moreno Uribe
University of Iowa

Brian Howe (✉ brian-howe@uiowa.edu)
University of Iowa

Research Article

Keywords: Dental Anomaly, Intraoral Photos, predictions

Posted Date: November 30th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1061414/v1

Abstract

Children with orofacial clefting (OFC) present with a wide range of dental anomalies. Identifying these anomalies is vital to understand their etiology and to discern the complex phenotypic spectrum of OFC. Such anomalies are currently identified using intra-oral exams by dentists, a costly and time-consuming process. We claim that automating the process of anomaly detection using deep neural networks (DNNs) could increase efficiency and provide reliable anomaly detection while potentially increasing the speed of research discovery. This study characterizes the use of` DNNs to identify dental anomalies by training a DNN model using intraoral photographs from the largest international cohort to date of children with nonsyndromic OFC and controls (OFC1). In this project, the intraoral images were submitted to a Convolutional Neural Network (CNN) model to perform multi-label multi-class classification of 10 dental anomalies. The network predicts whether an individual exhibits any of the 10 anomalies and is able to do so significantly faster than a human rater. For every anomaly except mammalons, F1 scores suggest that our model performs competitively at anomaly detection when compared to a dentist with 8 years of clinical experience. In addition, we use saliency maps to provide a post-hoc interpretation for our model's predictions. This enables dentists to examine and verify our model's predictions.

Introduction

Individuals with orofacial clefting (OFC) present with a wide range of complex dental anomalies that affect tooth size, shape, structure, number, symmetry, and position, thus increasing phenotypic complexity and dental morbidity in affected individuals. Amongst these anomalies, the most common ten types include hypoplasia, hyopcalcification, agenesis, mammalons, microdontia, supernumerary teeth, impacted teeth, tooth rotations, and displacements. Although dental anomalies may often appear in the general population (up to 22% in the primary and 47% in the permanent dentition), their occurrence in individuals affected with overt clefts is much higher (up to 45% in the primary and 61% in the permanent dentition) and their etiology remains unknown[1-6]. Accurate and efficient identification of dental anomalies is vital to understanding their etiology, management and prevention. Specifically, the development of methods for large-scale screening of dental anomalies in human populations with high accuracy and effectiveness will largely increase the precision of association or causality estimates of genetic and environmental effects on such anomalies. In this work, we identify inefficiencies in the screening process and propose a deep learning based method to address them.

Currently, in-person dental exams, review of radiographs, and/or intraoral photographs are used to identify and document dental anomalies. However, these methods are labor-intensive, requiring training and careful calibration and are very time consuming, particularly for large samples, and thus can in turn slow down the speed of discovery. For instance, in our previous studies with large data, it took 1 year for a human rater to score over 30,000 intraoral images (IOPs) in 4084 subjects[6]. In addition, in the current human rater method, bias and errors in identification can occur and thus inter and intra-rater reliabilities of the dental anomaly data acquired are important aspects of data integrity that must be considered. These challenges are compounded in multicenter studies since an increase in the number of raters is required to complete data collection efficiently. Machine learning methods such as deep learning may be a promising solution to score large data sets objectively, reliably, and efficiently. While it takes years to train a human rater, in only takes hours to train a machine learning model. We also claim that in the long run, using machine rather than human labor saves significant time in scoring and can increase discovery speed.

In recent years, convolutional neural networks (CNNs) have become a state-of-the-art solution for image classification and have been successfully applied to dentistry[9]. CNNs are a particular type of Deep Neural Network (DNN). A recent survey[9] reported on about 30 published papers (as of April 2021) in the intersection of deep learning and dentistry. Examples include using CNNs to detect periapical lesions, dental caries, and odontogenic cystic lesions. However, it indicates that only very few publications[7,8] use digital camera photos as input data. The majority of existing work trained a CNN model using medical images such as radiographs or computed tomography scans that must be obtained by medical devices and are costly for patients. Our paper differentiates from previous CNN methods applied to dentistry in the use of digital cameras as opposed to specialized equipment, which both simplifies the data collection process and saves on hardware costs. The improved obtainability of the image makes our method more accessible for patients and easier to use.

A potential challenge for deep learning is that, in order to perform well, deep learning methods rely heavily on the amount of available training data. The models need to “see” enough examples to fit the large number of parameters. Previous dental literature used relatively small data sets of, at most, a few thousand images[7-9]. Our data set presents a unique opportunity to implement a deep learning method by having access to a sample that is orders of magnitude larger than previous research, collected from the largest international cohort, to-date, of subjects with OFC and controls.

Besides having a relatively large training dataset, our model also benefits from transfer learning (TL). The technique of TL starts with acquiring a trained CNN image classifier developed using a large number of images. The next step is to re-train the classifier on a new dataset but usually with the weights of the first few layers kept unchanged (frozen). Transfer learning can improve the predictive performance of a CNN because the low and mid-level feature transformation is very similar across different image classification tasks regardless of the target variable. Thus, our model can effectively “borrow” knowledge from existing state-of-the-art models. The TL technique can mitigate the lack of data problem as it uses information from other sources to build the model. We have identified only 13 publications in the dental image classification literature since 2017 that have utilized TL[7,8,10-18].

While deep learning models can achieve highly accurate predictive performance, their “black-box” nature has been criticized for hindering human understanding[29], especially in medical applications. Therefore, in addition to classifying the presence of anomalies, we also aim to provide interpretable post-hoc explanations for why the model makes such a diagnosis, by showing users which part of the image the model has focused on for a given intraoral photo. To do that, we generate a saliency map highlighting the area that is considered most important for the CNNs output. Fig. 1 shows the workflow of our data analysis.

In summary, the goal of the present study is to use deep learning (CNN and TL) to classify dental anomalies (agenesis, hypoplasia, hypocalcification, impacted teeth, incisal fissures, mammelons, microdontia, supernumerary teeth, and tooth rotation and displacement) for an input IOP and be comparable yet highly efficient compared to an expert human rater. To enable human interpretation, we generate saliency maps to provide explanations for how the CNN classifies images as having or not having dental anomalies, allowing verification of the predictions when using our method in practice.

The ability to do image classification with intraoral photos is foundational to facilitating personalized dentistry. In such a system, the practitioner would be able to upload intraoral photos and get diagnostic results via a trained neural network as well as explanations for the anomaly identification via a saliency map, thereby aiding the dentist in their diagnostic abilities and expediting clinical visits. Automating the process of dental anomaly classification using DNNs could also increase reliability, reproducibility, and the speed of anomaly classification. If a DNN is equal to or superior to human-raters while being significantly faster and more objective, the DNN-based image processing framework has the potential to revolutionize data collection methods and increase the speed of research discovery in orofacial biology and beyond.

Methods

Dataset

The study consisted of 38,486 intraoral photographs in 4,084 subjects (765 with OFC and 3319 control subjects). Intraoral photos and associated anomaly data was utilized from a previous study (OFC1)[6] from multiple sites in the United States and internationally. This study was reviewed by the Internal review board (IRB) at the University of Iowa and determined to be exempt from IRB review. All methods were carried out in accordance with relevant guidelines and regulations. Informed consent was obtained from each subject or their legal guardian(s) as part of the original study (OFC1). For each subject, a series of 6-10 intraoral photographs were made to fully display the entire oral cavity. Information corresponding to each subject, such as cleft status and the presence of various anomalies on each tooth, was logged into the OFC1 database. Each subject's photo set was evaluated and scored for dental anomalies using a paper form developed for this use (Supplementary Fig. S1). The rater, BJH, was a dentist with 8 years of clinical experience and was calibrated against two more experienced dentists for identification of dental anomalies on a small validation dataset prior to OFC1 data collection. Intra-rater reliability for BJH was 100% agreement with kappa = 0.95. Inter-rater reliability between all three raters was between 97.1-97.3% agreement with kappa= 0.91-0.93. After calibration, BJH scored all subjects and photos, becoming the ground truth. The training data for our CNN are constructed from OFC1 as follows. Given a photo, we assign the label for that photo a length 10 binary vector where each of the 10 indices corresponds to one of the 10 anomaly types we consider. The vector is 1 at an index if the patient in the photo has the corresponding anomaly *on any tooth*, and 0 otherwise. For additional details used in the collection of data, see the supplementary material.

Model Architecture

We adopted state-of-the-art methods for image classification by using transfer learning with a popular CNN architecture. There are a few classic resources in which details on CNNs may be found[19]. In addition to a CNN, we adopted TL[20], which utilizes a pre-existing CNN that has been trained on a very large dataset of photos and adapted it to our task. This allows us to further boost the predictive performance and save a substantial amount of computation time.

The pre-trained CNN we have chosen is ResNet-18[21]. It is an 18-layer CNN that has been trained using fourteen million images from the ImageNet database[22]. We experimented with freezing a different number of layers while leaving the rest of the layers trainable to adapt the model to our dental anomaly classification task. Results show that the best performance was achieved when freezing the first 7 layers. Since each study subject can have multiple anomalies, we designed a multi-label multi-class output layer with 10 nodes, each representing a type of anomaly. Each node then produces a probability for an input to have the corresponding anomaly. Our model uses raw pixel data from intraoral photos and preprocesses them using the standard ImageNet procedure. For additional information on network architecture and methods, see the supplementary material.

Training and Evaluation

We tasked our CNN with making accurate classifications of dental anomaly presence in each photo, judging it by means of accuracy, F1, ROC/AUC, and precision/recal metrics. The dataset used to train, test and validate the model consists of the 38,486 photos in OFC1. We conducted a group 5-fold cross validation of our model. This cross-validation variant splits the data into five subsets such that each subset consists of 20% of the data. It differs from standard cross validation in that it splits data by patients, which ensures that patients are not represented in more than one fold, and each fold represents approximately the same number of patients. In each fold, four subsets of them are combined into a training set while the remaining is the testing set. This is done five times such that each subset is used as a test set once. We set the batch size to be 512 images, number of epochs to be 1000, and the initial learning rate to be 1.34E-6. We use the AMSGrad[23] variant of the AdamW optimizer in PyTorch[24]. For each epoch, the model takes in a batch of images and uses the AdamW optimizer to optimize the parameters in the fully connected layer to minimize the multi-class dice loss between the outputs and the true values. We found this loss function to yield better results than other means of tackling class imbalance, including using weighted binary cross entropy loss or focal loss. Using the principle of early stopping, if the model sees that in 60 consecutive epochs the validation loss has not decreased, it will cease training early to prevent overfitting. Further details appear in the supplementary material.

Saliency Maps

To provide an interpretable explanation to the results provided by our CNN, we generated a saliency map for each output, to show what regions of an input image were considered important by the model to produce the corresponding classification[25]. One may consider the outputs of a CNN as a vector of differentiable probability functions. A saliency map is a heat-map where the intensity of each pixel is calculated by taking the gradient of the functions produced by the CNN for each of the anomalies. The value represents the contribution of the corresponding pixel of an input image to a class score[26]. The higher the value, the more important the pixel is for the CNN model's classification decision. These gradients are computed per color channel of the input image. To obtain a heat map of gradients across an image, the max gradient can be used over each color channel. In our max gradient saliency maps, the color of each pixel ranges from blue (cold) to red (hot) depending on how big the max gradient was for that pixel. The saliency map allows for interpretability of the image and confirms that the CNN model is reliably identifying the correct anomalies.

Results

Predictive Performance

We evaluated our model using the test sets of each of the five folds. We report F1, ROC/AUC, precision, and sensitivity for each anomaly for our model in Table 1. For our model, the median F1 score from the 5-folds for each anomaly, which is a reflection of the specificity and precision of the model, ranged from 0.352 to 0.984, with rotation having the highest F1 score and hypoplasia having lowest F1 scores (0.989 and 0.352 respectively). The median AUC for each anomaly ranged from 0.683 to 0.872 with displaced teeth having a lowest AUC (0.66). The model had special difficulty in classifying incisal fissures and hypoplasia. The frequency of anomalies per image can be found in Supplementary Table S1.

Anomaly	F1	Precision	Recall	ROC AUC
Mammalons	0.584±0.035	0.563±0.026	0.654±0.051	0.845±0.008
Impacted	0.554±0.146	0.529±0.144	0.586±0.015	0.866±0.078
Hypoplasia	0.370±0.052	0.320±0.074	0.479±0.152	0.752±0.020
Incisal Fissure	0.394±0.116	0.366±0.100	0.432±0.148	0.787±0.068
Hypocalcification	0.741±0.008	0.620±0.014	0.922±0.016	0.719±0.018
Displaced	0.745±0.007	0.610±0.005	0.957±0.023	0.682±0.004
Microdontia	0.444±0.074	0.452±0.080	0.443±0.099	0.818±0.052
Supernumerary	0.466±0.093	0.446±0.107	0.493±0.085	0.845±0.070
Rotation	0.984±0.004	0.968±0.007	0.999±0.000	0.738±0.038
Agenesis	0.489±0.127	0.521±0.141	0.463±0.119	0.820±0.042

Table 1. Model Metrics. Results given are the mean result of all 5 folds with the standard deviation.

Comparison of CNN with Human Baseline

In addition to the above evaluate, we compare our model against a human rater. On a subset of 30 patients from OFC1, we record BJH's *pre-calibration* performance for the tasks of detection of each anomaly in Table 2. (Note that the data used to train and evaluate the model were labeled after BJH was calibrated) BJH classified whether or not each individual had each anomaly by examining all of their IOPs (this differs from our model, which classifies anomaly presence in each photo separately). LMU, a more experienced dentist, also classified the anomaly presence in the 30 patients. We used LMU's results as a ground truth to evaluate BJH's pre-calibration F1, precision, recall, sensitivity, and specificity metrics for each anomaly. F1 scores in Table 2 are recorded as 0 if BJH make no correct predictions. They are recorded as N/A if there were neither positive ground truth labels nor predictions of the positive label. Incisal fissures has a precision of N/A because BJH had neither true positives nor false positives. Supernumerary has a recall of N/A because BJH had neither true positives nor false negatives.

Anomaly	F1	Precision	Recall
Mammalons	0.857	1.000	0.750
Impacted	N/A	0.000	0.000
Hypoplasia	0.667	0.500	1.000
Incisal Fissure	0.000	N/A	0.000
Hypocalcification	0.400	1.000	0.250
Displaced	0.246	0.750	0.750
Microdontia	N/A	0.000	0.000
Supernumerary	0.000	0.000	N/A
Rotation	0.963	1.000	0.929
Agenesis	0.000	0.000	0.000

Table 2. BJH Pre-calibration Metrics

We use BJH's pre-calibration performance against LMU to get an idea of how our model compares with an actual dentist. We find that our model compares favorably to BJH. Although BJH's F1 scores for mammalons (0.857) and hypoplasia (0.667) are higher than the model's (0.584 and 0.37 respectively), BJH's F1 scores are lower for the remaining anomalies. See tables 1 and 2. We also found the difference in time required, on average, to classify anomaly presence to be significant. In this study, the training routines generally took on the order of 12 hours, while BJH has accumulated experience over 8 years of clinical experience. The test step took approximately 3 minutes for 7,697 photos, a rate of approximately 40 photos per second. Thus, if the model were to classify all 38,486 photos, it will need approximately 16 minutes to complete the task whereas it took a human-rater one year[6].

Post-hoc Interpretability via Saliency

To enable human understanding, we generated saliency maps to show important image regions when our CNN (correctly) predicts each of the 10 considered anomalies. See Fig. 2 for examples. For example, in Fig. 2a the saliency map highlights the incisal edge of the mandibular incisors, indicating that the CNN is recognizing the relevant area where mammalons occur and in Fig. 2b reveals hypocalcification on the maxillary right canine and the CNN highlighted the incisal edge areas. In addition, we also examined saliency maps for *incorrect* predictions, which is particularly important since if domain experts understand why the model makes a mistake, then they know when not to trust a model. We found that when a model makes a mistake, it often looks at non-relevant area of the images such as gingiva, buccal mucosa, or space between teeth. We also found that orthodontic appliances such as arch wires, brackets, and fixed retainers, are difficult for the CNN to ignore and could mislead the CNN. Orthodontic appliances can obscure dental anomalies for the CNN and human rater alike, thus this limitation could be applied to both. We also found that the CNN has difficulty with blurry or unfocused intraoral photos or those that depict a narrow or small field of view. We randomly sampled 100 mis-classified samples (10 for each anomaly type) and found that 21 had braces in them. 34 of them showed only a narrow region of the mouth. 4 of them were completely blurry. See Supplementary Fig. S3 for examples of saliency maps where the highly activated regions do not correspond with the actual locations of the anomalies.

Discussion

The use of image classification algorithms such as TL with CNNs has become increasingly popular in the past few years. Our findings suggest great potential in use of CNN-based image classification for quickly identifying dental anomalies from intraoral photos. The ability to produce saliency maps makes our method interpretable and provides insight into the model’s reasoning. Our method not only performs dental anomaly classification but can also show where in the mouth the CNN “looks” to make its decision. Clinicians and researchers can, therefore, consult the saliency map and verify whether the CNN model is making classifications that are consistent with the location and development of such anomalies. This can give additional confidence for clinicians and researchers using this model and can provide educational benefits for students and less experienced clinicians. In addition, since our method can work with intraoral photos taken by standard cameras, it is more accessible than other DNN based models that work with X-rays or CT-scans.

In the current study we used ResNet-18 as the pre-trained CNN model[21] for TL. This CNN was chosen over other models due to its low runtime and high accuracy when compared to other popular architectures on the ImageNet benchmark. ResNet-18 is a popular open-source network architecture, so theoretically if independent clinics were using our training methodology with separate private datasets, they could share model weights or training gradients in order to benefit from each-other’s data without sharing their data.

The dataset was originally scored for dental anomalies, by one person after calibration6 (also supplementary material) and took approximately one year of full-time work to score all 4,084 subjects and their respective 38,486 intraoral images. In the current study, the CNN would be able to identify the dental anomalies in the same number of photographs in approximately 16 minutes with F1 scores ranging from 0.32-0.989. Our results suggest that our model is able to perform at a similar level as that of a dentist with 8 years of clinical experience in the anomaly detection tasks. See additional metrics in Table 1. We found examples of both classification agreement and disagreement, for example where the model correctly predicted hypoplasia while the human rater did not, see Supplementary Fig. S2. This highlights the error that can occur from eye fatigue or human error that does not occur in computers.

To be successful at image classification tasks, a CNN needs to be trained on a very large number of examples in order to learn good feature representations from the input images. The size of the training examples has a direct impact on the overall model accuracy. The current data set is the largest international cohorts of intraoral photos of controls and subjects with OFC, with 38,486 images. For multi-label multi-class image classification task, this is still considered small. However, we were able to achieve moderate to high F1 scores (0.32-0.989) using the technique of transfer learning. To continue to improve and test the accuracy of this model, additional intraoral photographs will be needed. A second intraoral data set has been scored and will be used to further test and improve this algorithm to see if it can equal or outperform human raters on every dental anomaly. We used a separate sample of data to get an estimate of human performance with respect to the F1, precision, and recall metrics. For all but one type of anomaly, our model's F1 scores exceeded those of the human baseline.

Data imbalance between subjects with OFC and controls is a limitation of this study as subjects with OFC have a higher incidence of certain dental anomalies. To address this, we tested different loss functions that are supposed to be robust to data imbalance. We tested weighted binary cross entropy, multi-class dice loss, and focal loss. The multi-class dice loss proved to yield the best performance, in accordance. Another limitation of the current algorithm is that it does not give dental anomaly data per tooth, but whether any of the anomalies are present in the photograph per subject. Future work is needed, and is currently underway, for the CNN to identify each tooth in each photo and the associated anomalies.

In examining the saliency maps generated by the model, we found that orthodontic appliances such as arch wires, brackets, and fixed retainers, are difficult for the CNN to ignore and is a limitation of the study. Orthodontic appliances can obscure dental anomalies for the CNN and human rater alike, thus it is a limitation for providers and the CNN. Blurry or unfocused intraoral photos or those that depict a narrow or small field of view are also a limitation of this study. This limitation can be solved by providing more high quality photos to the model.

This algorithm also has the potential to be a second rater to calibrate against or even a replacement for the rater with further validation, which will increase the speed of data collection and analysis while saving cost. This method could be used in the field when intraoral-photos are made, uploaded, run through the algorithm and the results transmitted to the principal investigator from sites around the world, thus the person-hours needed for dental anomaly classification could decrease significantly assisting oral health research around the globe. Another possible application would be a dental phenotype-to-gene or tooth-to-gene, where the CNN identifies the dental anomalies per subject and link this with an available genetic database to produce possible genes linked to the identified dental anomalies, similar to FACE2GENE (FDNA, Boston, MA).

Conclusion

In this work we proposed to use ResNet-18 and transfer learning to detect the presence of 10 dental anomalies using Intra-Oral Photos (IOPs) from standard cameras as inputs. In isolation, we found our method to obtain fairly good classification accuracy. When compared to human dentists, our method boasts significantly faster classification speed and competitive accuracy. To mimic the way human dentists can point out where they looked to recognize the presence of a dental anomaly, we used saliency maps to show where our model was looking when making predictions, which enable human dentists to understand the reasoning of our model.

Our algorithm has the potential to change how dental anomalies are scored and thus how dental anomaly phenotypes are identified in populations. It can greatly increase the speed of discovery by taking a task that potentially can take years, with a large data set similar to the current one, to taking a couple of hours. Using it instead of or in tandem with human raters would lower long-term costs for identification of dental anomalies. In the future, for image analysis of dental anomalies, data collection and analysis may take place simultaneously, transmitted to the research team for the findings to be interpreted via a secure website, which is under development. Further research is needed in this exciting area of dental research.

Declarations

Data and code availability

The code used in this study is available here https://github.com/rrags/DentalAnomalyDetector. Data is available upon request for mutual collaboration.

Acknowledgements

A special thanks to all the families that participated in this study. Also special thanks to Joel Anderton, Jennifer Jacobs, Beth Emanuele, and Carla Sanchez for their logistical support. Grant support was provided by: K08 DE028012: “Identifying Phenomic Patterns of Dental Anomalies in Orofacial Clefting” (University of Iowa as Primary Awardee). R01 DE106148: “Extending the Phenotype of Nonsyndromic Orofacial Clefts” (University of Pittsburgh as Primary Awardee). R01 DE014667: “Cleft Lip Genetics: A Multicenter International Consortium” (University of Iowa as Primary Awardee). R37-DE-08559: “Molecular Genetic Epidemiology of Cleft Lip and Palate” (University of Iowa as Primary Awardee). R01 DD000295: “Health Outcomes and Improved Phenotypic Characterization of Cleft Lip and Palate” (University of Iowa as Primary Awardee). R00 DE024571: “Genetic and Environmental Risks for Oral Clefts in Puerto Rican Hispanics” (University of Puerto Rico as primary awardee). S21 MD001830: “Hispanics in Research

Capability: SOHP and SOM Partnership" (University of Puerto Rico as primary awardee). U54 MD007587 "Puerto Rico Clinical and Translational Research Consortium" (University of Puerto Rico as primary awardee). R00 DE022378: Genetic Studies of Non-Syndromic Clefts in Populations of African Descent (University of Iowa as Primary Awardee). R01 DE02830: Refining the Genetic and Genomic Architecture of Non-Syndromic Orofacial Clefting (University of Iowa as Primary Awardee). Robert Wood Johnson Foundation 72429: Genome-Wide Association Studies for Non-Syndromic Clefts in sub-Saharan African Populations. The authors declare no potential conflicts of interest with respect to authorship and/or publication of this article.

Author contributions

A.B., C. P., J.T.H., F.A.P., I.M.O., C.J.B., C.V-R., C.R.M., and G.L.W. contributed to conception, design, and data acquisition, drafted and critically revised the manuscript; R.R. and T.W., contributed to conception, design, analysis, and data interpretation, drafted and critically revised the manuscript; B.J.H., S.M.W., M.L.M., and L.M.M-U. contributed to conception, design, data acquisition, analysis, and interpretation, drafted and critically revised the manuscript. All authors gave final approval and agree to be accountable for all aspects of the work.

Competing interests

The authors declare no competing interests.

Additional Information

Supplementary Information Supplementary material is available at <link>

Correspondence and requests for data should be addressed to B.J.H.

References

1. Eerens, K. *et al.* Hypodontia and Tooth Formation in Groups of Children with Cleft, Siblings without Cleft, and Nonrelated Controls. *The Cleft Palate-Craniofacial Journal* **38,** 374–378 (2001).

2. Letra, A., Menezes, R., Granjeiro, J. & Vieira, A. Defining Subphenotypes for Oral Clefts Based on Dental Development. *Journal of Dental Research* **86,** 986–991 (2007).

3. Rawashdeh, M. A. & Sirdaneh, E. O. A. Crown Morphologic Abnormalities in the Permanent. *Journal of Craniofacial Surgery* **20,** 465–470 (2009). 5. Wu T-T, Chen PKT, Lo L-J, Cheng M-C, Ko EW-C. 2011. The characteristics and distribution of dental anomalies in patients with cleft. Chang Gung Med J. 34(3):306–314.

4. Walker SC, Mattick CR, Hobson RS, Steen IN. 2009. Abnormal tooth size and morphology in subjects with cleft lip and/or palate in the north of England. Eur J Orthod. 31(1):68–75. doi:10.1093/ejo/cjn073.

5. Wu T-T, Chen PKT, Lo L-J, Cheng M-C, Ko EW-C. 2011. The characteristics and distribution of dental anomalies in patients with cleft. Chang Gung Med J. 34(3):306–314.

6. Howe, B. *et al.* Spectrum of Dental Phenotypes in Nonsyndromic Orofacial Clefting. *Journal of Dental Research* **94,** 905–912 (2015).

7. You, W., Hao, A., Li, S., Wang, Y. & Xia, B. Deep learning-based dental plaque detection on primary teeth: A comparison with clinical assessments. (2020). doi:10.21203/rs.2.21027/v2

8. Takahashi, T., Nozaki, K., Gonda, T. *et al.* Deep learning-based detection of dental prostheses and restorations. *Sci Rep* **11,** 1960 (2021). https://doi.org/10.1038/s41598-021-81202-x

9. Ren, R., Luo, H., Su, C., Yao, Y. & Liao, W. Machine learning in dental, oral and craniofacial imaging: a review of recent progress. *PeYou, W., Hao, A., Li, S., Wang, Y. & Xia, B. Deep learning-based dental plaque detection on primary teeth: A comparison with clinical assessments. (2020). doi:10.21203/rs.2.21027/v2 erJ* **9,** (2021).

10. Aubreville, M. *et al.* Automatic Classification of Cancerous Tissue in Laserendomicroscopy Images of the Oral Cavity using Deep Learning. *Scientific Reports* **7,** (2017).

11. De Tobel, J., Radesh, P., Vandermeulen, D. & Thevissen, P. W. An automated technique to stage lower third molar development on panoramic radiographs for age estimation: a pilot study. *Journal of Forensic Odonto-Stomatology* **35,** 42–54 (2017).

12. Murata, S., Lee, C., Tanikawa, C. & Date, S. Towards a Fully Automated Diagnostic System for Orthodontic Treatment in Dentistry. *2017 IEEE 13th International Conference on e-Science (e-Science)* (2017). doi:10.1109/escience.2017.12

11. Prajapati, S. A., Nagaraj, R. & Mitra, S. Classification of dental diseases using CNN and transfer learning. *2017 5th International Symposium on Computational and Business Intelligence (ISCBI)* (2017). doi:10.1109/iscbi.2017.8053547

12. Lee, J.-S. *et al.* Osteoporosis detection in panoramic radiographs using a deep convolutional neural network-based computer-assisted diagnosis system: a preliminary study. *Dentomaxillofacial Radiology* **48,** 20170344 (2019).

13. Lee, J.-H., Kim, D.-H., Jeong, S.-N. & Choi, S.-H. Diagnosis and prediction of periodontally compromised teeth using a deep learning-based convolutional neural network algorithm. *Journal of Periodontal & Implant Science* **48,** 114 (2018).

14. 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. *Journal of Dentistry* **77,** 106–111 (2018).

15. Zhang, K., Wu, J., Chen, H. & Lyu, P. An effective teeth recognition method using label tree with cascade network structure. *Computerized Medical Imaging and Graphics* **68,** 61–70 (2018).

16. Das, N., Hussain, E. & Mahanta, L. B. Automated classification of cells into multiple classes in epithelial tissue of oral squamous cell carcinoma using transfer learning and convolutional neural network. *Neural Networks* **128,** 47–60 (2020).

17. You, W., Hao, A., Li, S., Wang, Y. & Xia, B. Deep learning-based dental plaque detection on primary teeth: a comparison with clinical assessments. (2020). doi:10.21203/rs.2.21027/v2

18. Lin, H.-H. et al. On construction of transfer learning for facial symmetry assessment before and after orthognathic surgery. Computer Methods and Programs in Biomedicine 200, (2021).

19. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521,** 436–444 (2015).

20. Weiss, K., Khoshgoftaar, T. M. & Wang, D. A survey of transfer learning. Journal of Big Data 3, (2016).

21. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016). doi:10.1109/cvpr.2016.90

23. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (2009). doi:10.1109/cvpr.2009.5206848

24. Reddi, S. J., Kale, S. & Kumar, S. International Conference on Learning Representations. in *On the Convergence of Adam and Beyond* (2019).

25. Paszke, A. *et al.* Neural Information Processing Systems. in *PyTorch: An Imperative Style, High-Performance Deep Learning Library* (Curran Associates, Inc., 2019).

26. Ogura, M. & Jain, R. *FlashTorch* (2020).

27. Simonyan, K., Vedaldi, A. & Zisserman, A. International Conference on Learning Representations. in *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps* (2014).

28. Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S. & Cardoso, M. J. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support Lecture Notes in Computer Science* 240–248 (2017). doi:10.1007/978-3-319-67558-9_28

29. Rudin, C. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nature Machine Intelligence* 1.5 (2019): 206-215.

Figures

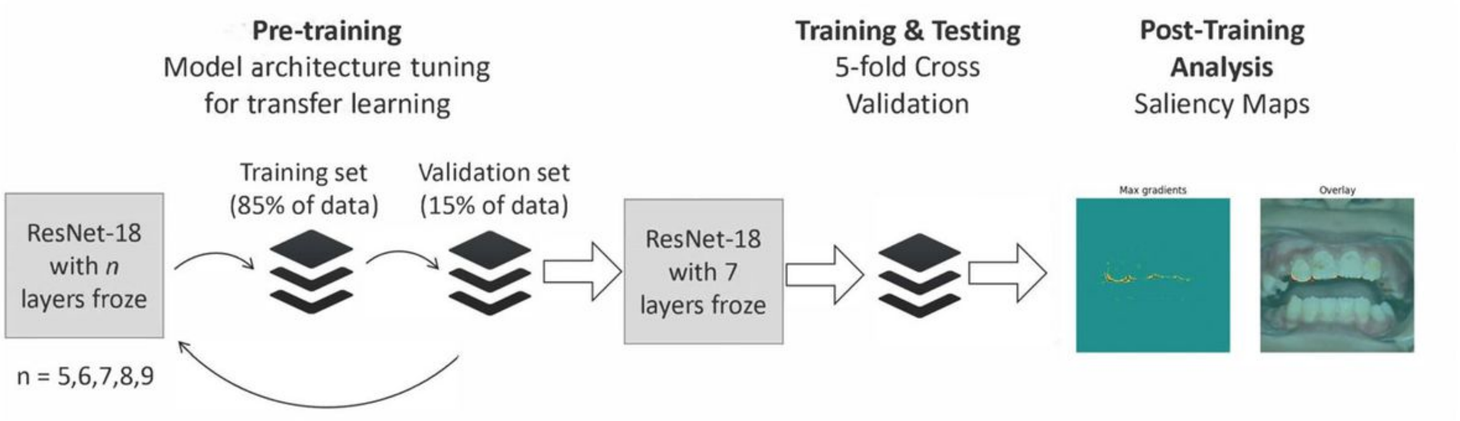

Figure 1

Workflow of data analysis. The entire workflow consists of three steps. In step 1, we tune the number of layers to freeze in order to do TL optimally. Our experiments show that when freezing 7 layers, our model achieved the best predictive performance. We then test the model using a 5-fold grouped cross-validation. Finally, for each input photo and the corresponding model prediction, we generate a saliency map for each anomaly (regardless of presence in the photo).

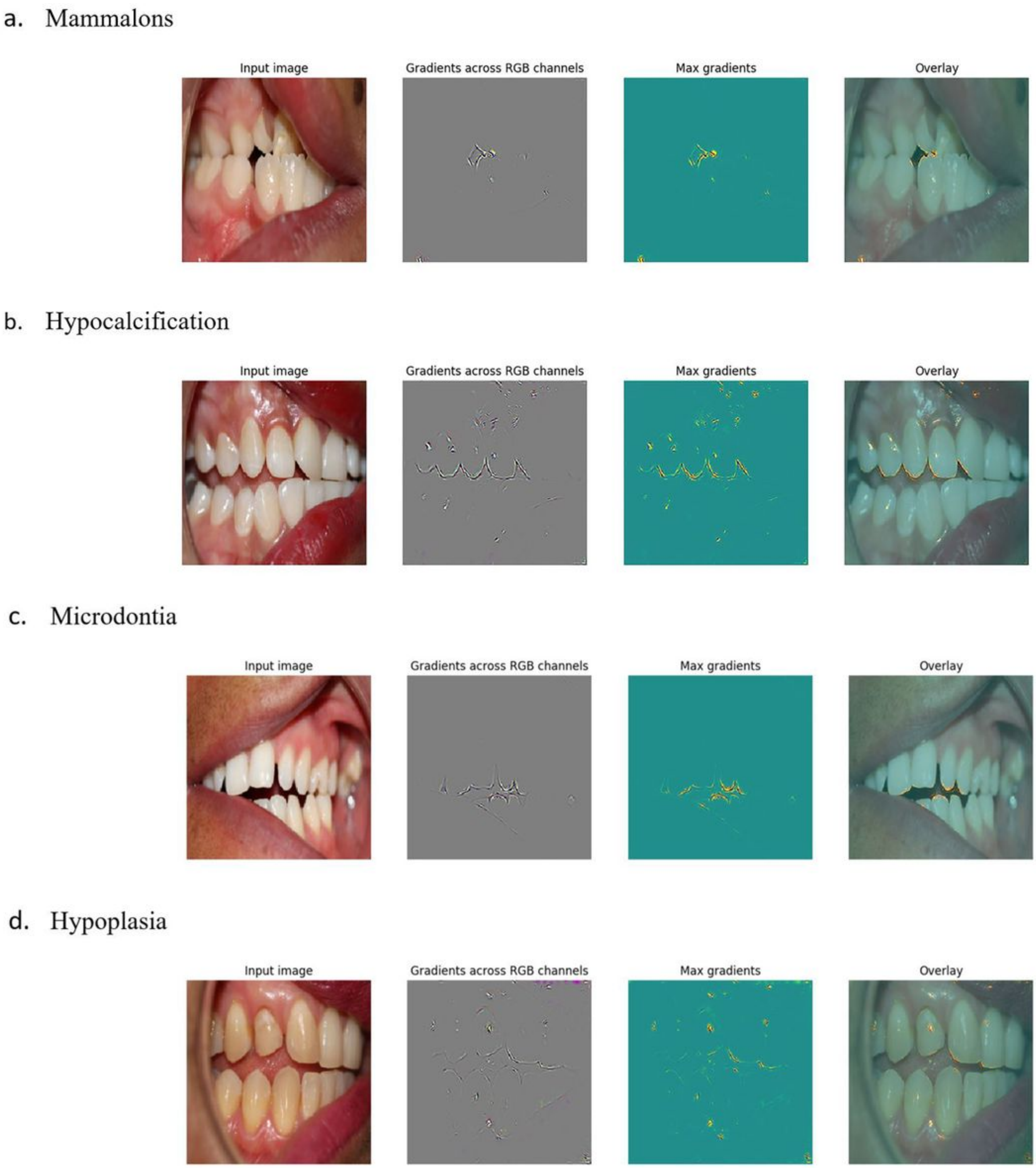

Figure 2

Saliency Maps. Note: Overlay is the input image overlaid with the gradients. These are representative examples of anomalies depicting what the algorithm saw when making correct predictions of mammalons, hypocalcification, microdontia, and hypoplasia.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterialNSR11192021.docx