2018 42nd IEEE International Conference on Computer Software & Applications

Automated Dental Image Analysis by Deep Learning on Small Dataset

Jie Yang[1], Yuchen Xie[1], Lin Liu[1], Bin Xia[2], Zhanqiang Cao[2], and Chuanbin Guo[2]

[1]Tsinghua University
[2]Peking University Stomatological Hospital
{yangj16,xieyc17}@mails.tsinghua.edu.cn,linliu@tsinghua.edu.cn
{xiabin,caozhanqiang}@pkuss.bjmu.edu.cn

***Abstract*—Dental radiography provides important evidence for clinical diagnosis, treatment and quality evaluation. Much effort has been spent on developing digitalized dental X-ray image analysis systems for clinical quality improvement. In this paper, we present the datasets, procedures, and results conducted to evaluate dental treatment qualities using periapical dental X-ray images taken before and after the operations. In order to support dentists to make clinical decisions, we propose a tool pipeline for automated clinical quality evaluation. We build a dataset with 196 patients' periapical dental radiography images before and after the treatments. Radiography images are labelled as cases that are 'getting better', 'getting worse' and 'have no explicit change' by designated dental experts. Our proposal includes an automatic method with the medical knowledge to crop the ROIs for clinical evaluation – the apical adjacent regions, and then pairs of ROIs are fed into a CNN to train the model for automated clinical quality evaluation. Our approach achieves the F1 score of 0.749, which is comparable to the performance of expert dentists and radiologists.**

***Index Terms*—Deep Learning, Medical Image Processing, Convolution Neural Network (CNN), Clinical Decision Support (CDS)**

I. Introduction

Dental clinical quality evaluation includes many quantitative measures, such as the statistics of annual dental visits of children or adults of different ages, new caries among caries-active children, endodontics to extractions procedure ratio [1], etc. While these quality measurements evaluate dental clinical quality at the macro-level, we also need to conduct qualitative evaluations at the micro-level, such as to understand whether the treatment conducted to an individual tooth take effect or not. For example, whether the root canal therapy conducted has prevented further infection of chronic endodontics and periapical diseases effectively. Conventionally, oral healthcare experts evaluate the patients' dental conditions and dentists' treatment qualities by observing changes between the dental images taken prior to and after the treatment.

In this paper, we propose an automated root canal treatment quality evaluation procedure based on dental radiograph image classification, which integrates medical experts experience, with image processing algorithms and convolutional neural network-based learning. It is part of an oral clinical data analytics project, a joint effort of five Schools of Stomatology, one IT research lab and four Health-IT Corporations partners. The research question was brought up by a head pediatrician. Although experienced dentists only need a few seconds to make evaluation to one set, the workload will become very heavy if there are hundreds of cases, and sometimes mistakes will appear. On the other hand, the automated system can be used for interns learning platform for education purpose.

The experimental dataset includes 196 pair-wised periapical dental X ray images, which are categorized into three different groups: cases whose conditions 'getting better' significantly after the root canal treatment; cases whose conditions have 'no obvious changes' after the root canal treatment; cases whose conditions 'getting worse' after the root canal treatment. All the cases are labelled by experienced dentists.

As the time intervals of each pair of dental radiology examination vary from two weeks to two years, they could be taken by different radiologists, using slightly different radiation dose, angle, focal position, etc. Thus, the important visual features for clinical decisions might be mixed with noise pixels in the input images. In order to solve this problem, we develop an image registration process to unify each pair of images, so that the comparison results will be more accurate by only focusing on the ROIs rather than irrelevant pixel differences.

As the majority of input images in our study include root canal filling operation, and the filling materials are high density areas in dental pulp cavity. It is the most recognizable feature in post-treatment dental images, and also the area of the greatest pixel changes before and after treatment. In this paper, we first use image subtraction techniques to identify the root canal filling area. Then, we use the filling area to identify the tooth being treated to obtain the coordinates of the apical foreman and its adjacent area. We then feed this portion of the images to the CNN training program to train our classification models of various complexity. Considering the nature of the dental images, we select deep neural networks with few layers to refrain from over-fitting. Considering GoogLeNet includes too many layers to yield ideal results for our problem, we reconstruct its Inception structure to include two cascaded convolution layers to reduce the number of parameters and maintain a considerable performance of the network. Our experiment results show that the overall F1-score is 0.749, which is comparable to the average performance of expert-

0730-3157/18/$31.00 ©2018 IEEE
DOI 10.1109/COMPSAC.2018.00076

level dental practitioners even with small training dataset.

We have three major contributions to the medical image processing community:

- We propose an automated apical foreman-area identification approach from dental images for root canal filling treatment;
- We build a dataset of 196 pairs of pre-treatment and post-treatment dental images for clinical quality evaluation of root canal filling treatment;
- We conduct a comparative study to different convolution network structures and evaluate their feasibility and performance to the given clinical quality evaluation problem based on image classification.

II. Related work

Lots of researches are focused on image segmentation and classification. Both traditional digital image progressing algorithms and deep learning methods are applied in these fields.

In image segmentation area, traditional image segmentation methods are widely used in medical images, including region based methods [2], edge tracing methods [3] and integration methods [4]. The above methods are more suitable for specific problems, e.g. atlas-based segmentation are suitable for MR imaging sequences segmentation [5]. Applying deep learning to medical image segmentation is a common subject recently, and as such has also seen the widest variety in methodology, including the development of unique CNN-based segmentation architectures and the wider application of RNNs [6] [7].

Tooth canal edge extraction forms a pre-requisite step prior to many other advanced automated processing of dental X-ray images. However, data variation and fuzzy contours in dental radiography makes it a difficult task. Gayathri et al. [8] applied conventional edge extraction operators, Laplacian of Gaussian (LOG) edge extraction, Canny edge extraction and Zero Frequency Resonator (ZFR) based method for edge extraction. But the extraction is often incomplete or insufficient as there are extra noise areas.

For the challenge of bitewing radiography segmentation in [9], they collected a dataset contains 120 bitewing radiography images and manually labelled seven types of tooth structures, such as enamel, dentin and pulp. Lee et al. built a random forest based dental segmentation system. The average F-score of the above teams are 0.560 and 0.268 respectively. Ronneberger et al. [10] presented a machine learning approach using a U-shaped deep convolutional neural network (U-net) for the fully automated segmentation of dental X-ray images and they used data augmentation by applying elastic deformations.

Since manual labelled in pixel-level are costly and erroneous for large datasets, we aim to develop an unsupervised method for tooth contour extraction.

As for medical image classification, there are many attempts to use deep learning, especially CNNs. Medical image classification studies have done in human brain, eye, chest, cancer and other medical effects [11]. The Google team used a 23-layer CNN to detect diabetic retinopathy [12]. Esteva et al. [13] used GoogLeNet Inception v3 for skin cancer and it outperforms the average of the dermatologists at skin cancer classification using photographic and dermoscopic images. CNNs are also used in arrhythmia detection. Several research studies focused on dental image recognition and classification, such as in caries detection [14], teeth classification [15].

In the small dataset image classification cases, methods based on digital image processing are used more widely. Computed Tomography (CT) texture features are used to make abnormal classification by KNN and SVM [16]. Pyramid Histogram Of visual Words (PHOW) and Pyramid HOG (PHOG) as features are used to train SVM [17]. Bag-of-features approach like SIFT are applied SVM as classifier [18].

In summary, there is no ready-made technical solutions for dental image diagnosis in the literature, but there are related research effort to each individual steps of image processing and analysis using digital image processing methods, conventional classification techniques and deep neural networks.

III. Research Methodology

A. Overview

Figure 1 shows the overall research procedure of our study. The first step is to obtain the experimental dataset. The dental experts provided a labelled dataset, which made up of 196 pairs of periapical radiography images taken before and after treatment. Our expert team includes three professional dentists and a chief radiologist. We extract the ROIs – the apical region in image preprocess procedure: We first use the SIFT and SURF algorithms to find feature points between the pair of images. Then we find the best ternary matching points using minimum greyscale difference method. After that we calculate the affine matrix based on the best ternary points and use the image after affine transformation to do image subtraction. We exploit canal filling regions, which is obvious in images, help us find the root region. Then we crop the root apical portion and merge the two images of pre- and post- treatment in 2 channels, finally we feed these preprocessed images into our designed 6-layer CNN to make the clinical quality evaluation.

B. Tooth apical portion extraction

Due to the limited amount of labelled data, training directly with original images does not yield very good results. So we reconsider the solution, decide to further process the input images for training the classification network.

Actually, a typical periapical dental radiography image contains three or four teeth with the one to be treated in center approximately. According to the dentists experience, they are more concerned about the apical area when evaluating the clinical outcomes. In other words, if the machine can automatically identify the apical region of the infected root canal, and ignore the rest of the images, the diagnosis results will be improved considerably. And filling region, which differs in pre- and post-treatment images, is more clear in subtraction. It can help us finding apical area. Thus, we decide to use the following method to deal with the pre-treatment and

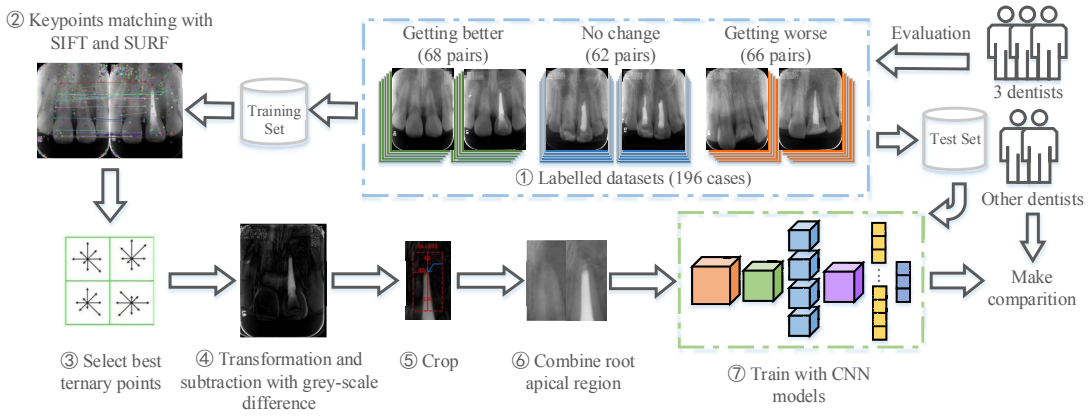
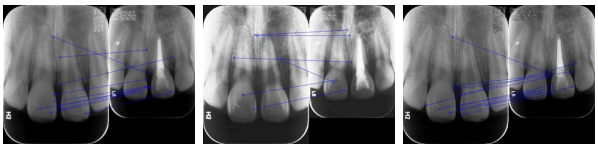


Fig. 1: Automated Analysis of Dental Images for Clinical Outcome Evaluation.



(a) original image (b) applying histogram equalization (c) applying USM

Fig. 2: Key points augmentation by combining histogram equalization and USM.

post-treatment image pairs. We cut the apical portion for the future evaluation training.

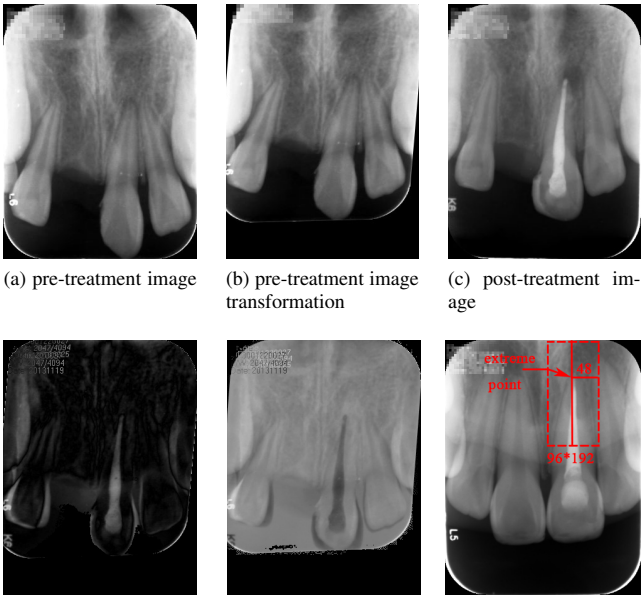
First, we use image calibration techniques to get the subtraction of the dental images taken before and after clinical treatment. The process includes two major steps:

1) Find the feature matching points using SIFT and SURF algorithm: SIFT and SURF algorithm are common image feature matching algorithms. However, the position, angle, light and other conditions changes increase the difficulties for finding features point pairs, we combine the two algorithms to get more key points to improve the accuracy.

For further improvement, we apply the histogram equalization method and unsharp masking method (USM) before finding feature points. Figure 2 shows key points found by SURF in three conditions: original image, applying histogram equalization and applying unsharp masking method. We combine all the key points found in these conditions for enhance key points.

2) Find the best ternary points according to the greyscale difference: In the automatic calibration, we believe that by treating the pre-treatment image translation, rotating and scaling, we can get a very similar result to the post-treatment image. Key points found in III-B(1) can be used for calibration.

Affine transformation is the mapping between two vector spaces. From an algebraic perspective, affine transformation can be described as matrix multiplication. If we get ternary points then we can calculate affine matrix and make auto



(a) pre-treatment image (b) pre-treatment image transformation (c) post-treatment image

(d) subtraction image (e) subtraction with retaining the derection of change (f) The specific crop range of apical region

Fig. 3: An example of subtraction and crop range.

image calibration. What we need to do is to find the best ternary points among key points.

We check all the ternary feature points and calculate their corresponding affine matrix, which are used to calibrate the images. We use both Random sample consensus (RANSAC) and minimum average greyscale difference methods. If the average greyscale difference of the calibrated images is minimum, we will consider the ternary points to be the most correct one.

3) Calculate affine matrix based on the best ternary points and use it to do image subtraction: We can calculate the corresponding affine matrix and obtain the calibrated output

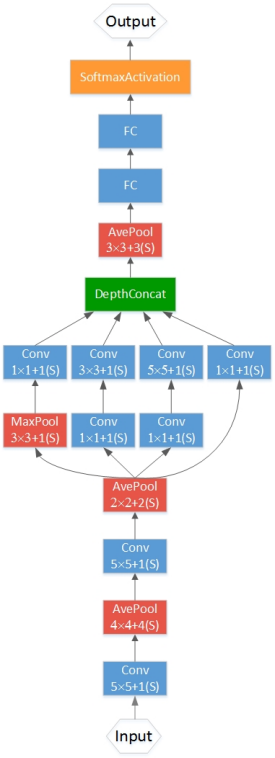


Fig. 4: The architecture of the deep neural network.

images based on the best ternary points. Figure 3 (a)-(e) gives an example of automated calibration. In subtraction image, filling region is more clear and easier to find. Comparing the two images, subtract the value of corresponding pixel and then we can get the change of the pixels. In this way, we can get the apical portion of root of the teeth being treated.

We use the disjoint-set data structure to find the adjacent apical area with the highest greyscale difference. The apical root is located on top or at the bottom of the filling area. We can determine whether it is an upper or lower tooth according to the greyscale distribution. Furthermore, we crop out the region within a certain range centered on this coordinate. The specific crop range is shown in the Figure 3 (f). We crop the region of 96 pixels $\times$ 192 pixels, the extreme point is not at the center of the region. (e.g. for upper incisors, we crop 130 pixels above the extreme point and 62 pixels below the extreme point). In this way, we include more adjacent space to the tooth apical for more accurate evaluation. For images which have more than more filling regions, adjust the threshold to cover it. To maintain better root canal filling region, we also adapt the erosion operation, over-exposure preprocessing and morphological testing.

C. Dental Image Classification Model

We use convolutional neural networks (CNNs) for clinical outcomes evaluation. Our current network structure is shown in Figure 4. Inputs of the network are the cropped images of periapical regions after the pre-processing procedure as described earlier in section III-B. We combine the preprocessed pre- and post-treatment images in 2 channels. Outputs are the classifications of the clinical quality. The training set and the test set are generated by random selection restricted by the proportion of 4 : 1, and images (including the source images and the transformed images from the same images) do not cross in the training set and the test set.

Our network is made up of 4 layers of convolutions followed by 2 fully connected layers and a softmax layer. First two convolutional layers are traditional layers and we use an Inception module followed. Due to limited dataset size, we design a network structure with fewer layers and parameters, and smaller convolution kernels as far as possible. We use kernels of size 1×1, 3×3, 5×5 rather than 7×7 to reduce parameters. The structure of Inception modules is also applied for dimension reduction. We use a concat layer to combine 2 images in 2 channels.

In order to prevent overfitting, we apply dropout between full connectivity layers and we add $L2$ norm into loss function. we use softmax to calculate the classification distribution after full connectivity. Our goal is to minimize polynomial logistic loss.

We use SGD to update the weights. Learning rate starts from 10^{-4}, and we drop learning rate in steps by multiply 0.1 in every 10^4 iteration. We set the momentum as 0.9.

IV. Experimental Results

A. Datasets

We identified and annotated a dataset of 196 pairs of periapical dental radiographs. Every radiograph contains one or more teeth for treatment and each pair constitutes of two radiographs taken pre- and post- treatment (Figure 3 (a), (c)). Experienced dentists label these teeth as ‘getting better’, ‘no change’, or ‘getting worse’ depending on clinical observation.

The camera settings for taking radiographs are: 70 KV tube voltage, 6 MA tube current, 0.20 Secs exposure time. During the examination, radiologists may make slight adjustment based on this setting. The shooting angle of each tooth is different, which is called bisecting angle technique.

In order to enrich the dataset, we perform a data augmentation procedure. In the process of radio-graphical examination, quality of images will be influenced by factors like exposure time, radiation angle, etc. We flip the image horizontally and vertically, make gradual rotations, and change the image brightness within a certain range to imitate different possible examination environment in real world.

B. Automated Calibration and Cropping of Apical Region

In the initial experiment, we use images include the entire tooth as input. However, cases like adjacent tooth overlap under the X-ray and tooth wear introduce noise to the problem. On the other hand, in order to avoid overfitting, we want to reduce the size of input images due to the limited sample cases in the dataset. According to the dentists’ experience, we extract the apical region.

classes	front-teeth		molars	
	calibration	crop	calibration	crop
getting better(68)	49/56	49/56	11/12	8/12
no cange(62)	29/33	27/33	20/29	15/29
getting worse(66)	31/35	28/35	14/31	9/31

TABLE I: Result of automated calibration and cropping.

We use the method mentioned in Section III-B for automated calibration and cropping of apical region on the entire dataset. Figure 3 shows an example of subtraction. Our method performs better than existing method [19] in fuzzy conditions.

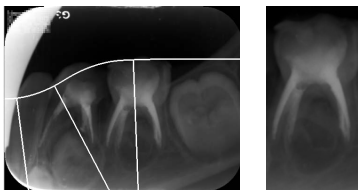
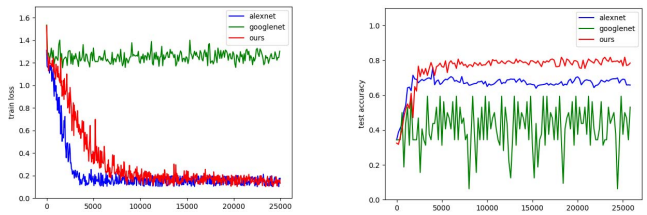


Fig. 5: Our method (right) in comparison with Chen et al. [19] (left) in fuzzy conditions.

Method	Histogram	subsampled	HOG	PCA	HOG+PCA
SVM	0.400	0.565	0.455	0.565	0.626
KNN	0.346	0.510	0.475	0.313	0.553
GBDT	0.471	**0.642**	0.475	**0.625**	**0.650**
RF	0.464	0.627	0.454	0.560	0.560
NB	0.347	0.580	0.440	0.327	0.577
AdaBoost	**0.497**	0.647	**0.499**	0.476	0.558

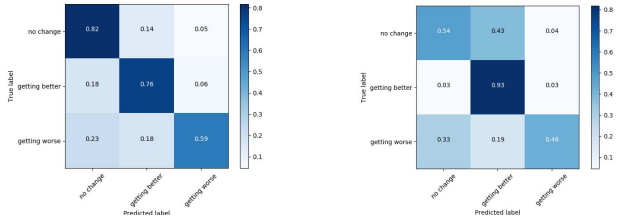
TABLE II: Result of F1 scores in baseline method.

As shown in Table I, the method of automatic calibration in this article yield good results on most dental images, especially on the front teeth. And for images of front teeth, once calibrated correctly, they will have better chance to get an ideal crop result. However, our method doesn't always work on images of molars. The differences between some images are so overwhelming that based only on affine matrix, even



(a) training loss (b) test accuracy

Fig. 6: Training loss and test accuracy.



(a) model (b) dentist

Fig. 7: Confusion matrix for the model predictions (left) and dentists labeling results (right) on the test set.

manual calibration also does not perform well. Therefore, we need to explore new methods to deal with these image pairs in the future.

C. Classification Network

We tried several network architectures for auto cropping apical portion dataset. As shown in Figure 6, our network structure has higher accuracy. It converges at 10000 iterations, which is slightly slower than AlexNet but achieve higher accuracy. And the deeper network GoogLeNet does not converge on our dataset. The proposed network yields best experimental performance, while adopting deeper network structure will lead to over-fitting.

D. Clinical Outcomes Evaluation

1) Baseline: We compare our approach with baseline methods on image classification (Table II). All of these baseline methods use feature extraction and a classifier. Features are extracted from manual cropped root apical region images.

We use Histogram of Oriented Gradients (HOG), Principal Components Analysis (PCA), Greyscale Histogram and subsampling method to extract image features, then we use Support Vector Machine (SVM), k-nearest neighbors (KNN), Gradient Boosting Decision Tree (GDBT), Random Forest(RF), Naive Bayes (NB) and AdaBoost to make classification. We tuned parameters of both features and classifiers to show the best F1 score in each algorithm.

As shown in Table II, classifiers like GBDT and AdaBoost achieve higher scores. Moreover, processing HOG features by PCA and apply GBDT achieves highest F1 score of 0.650.

2) our approach: Table III shows the average F1 scores of multiple experiments with the whole tooth and root apical areas as inputs respectively. We can see F1 scores are improved when using root apical region instead of the entire teeth. We try two ways to combine pre- and post- treatment images: combining them into one channel side-by-side, putting them in two channels, the results show the latter is better, which can also reduce parameters. The cases under 'getting better' category achieved best results, while cases under the 'getting worse' category achieved less good results.

We then invited two dental experts and one chief radiologist to participate in the comparative study. Each dentist labelled the test dataset separately. We find that there are obvious

Method	the whole tooth	manual crop	auto crop^{+}	auto crop*	dentist 1	dentist 2	dentist 3
Class-level F1 Score							
getting better	0.456	0.826	0.750	0.763	0.769	0.769	0.667
no change	0.578	0.742	0.718	0.754	0.444	0.706	0.727
getting worse	0.417	0.729	0.703	0.729	0.182	0.857	0.571
Aggregate Results							
Precision	0.537	0.767	0.742	0.756	0.681	0.828	0.746
Recall	0.490	0.778	0.729	0.742	0.461	0.785	0.667
F1	0.517	0.772	0.735	0.749	0.550	0.806	0.704

TABLE III: the F1 scores of different ways to make quality evaluation. auto crop^{+} combine two images side-by-side and auto crop* use two channels.

distinctions between different dentist results. Similar to our network, the average of dentists results achieve the highest F1 score in the 'getting better' cases, and lowest in the 'getting worse' cases. Moreover, during the labeling process, dentists can make decisions more quickly to cases that the tooth is 'getting better'. In contrast, 'getting worse' and 'no change' are less obvious. It is also shown in the confusion matrix (Figure 7), our proposed network and the dentists both achieve higher accuracy in the 'getting better' cases but more easily to make wrong decisions on the 'getting worse' cases.

From the Table III we can see that the F1 score of our network is higher than the average of results of dentists on the 3 classes, which confirms that our automated evaluation streamline approach performs equally well as experts.

V. Conclusion and Future Work

We have developed an automated, streamlined dental image analysis approach that integrates dental image diagnosis knowledge. It supports the automated apical region identification which saves much manual efforts on data preparation. Our approach supports CNNs for diagnosis classification based on small datasets. With limited labelled cases, our approach yields promising results that is comparable to expert-level dentists and dental radiologists. In particular, our method achieves considerable results when it comes to identifying 'getting better'. Our study also confirms that for medical image analysis problems, domain knowledge has to be considered when designing the data analytic steps and interpret the analysis results.

References

[1] D. Q. A. (DQA), "Pediatric oral health quality and performance measures: Environmental scan, 2012."

[2] S. Hojjatoleslami and F. Kruggel, "Segmentation of large brain lesions," *IEEE Transactions on Medical Imaging*, vol. 20, no. 7, pp. 666–669, 2001.

[3] A. Martelli, "An application of heuristic search methods to edge and contour detection," *Communications of the ACM*, vol. 19, no. 2, pp. 73–83, 1976.

[4] A. Chakraborty, M. Worring, and J. S. Duncan, "On multi-feature integration for deformable boundary finding," in *Computer Vision, 1995. Proceedings., Fifth International Conference on*. IEEE, 1995, pp. 846–851.

[5] F. Ségonne, A. M. Dale, E. Busa, M. Glessner, D. Salat, H. K. Hahn, and B. Fischl, "A hybrid approach to the skull stripping problem in mri," *Neuroimage*, vol. 22, no. 3, pp. 1060–1075, 2004.

[6] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *3D Vision (3DV), 2016 Fourth International Conference on*. IEEE, 2016, pp. 565–571.

[7] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. Springer, 2016, pp. 179–187.

[8] V. Gayathri, H. P. Menon, and A. Viswa, "Challenges in edge extraction of dental x-ray images using image processing algorithms–a review," 2014.

[9] C.-W. Wang, C.-T. Huang, J.-H. Lee, C.-H. Li, S.-W. Chang, M.-J. Siao, T.-M. Lai, B. Ibragimov, T. Vrtovec, O. Ronneberger *et al.*, "A benchmark for comparison of dental radiography analysis algorithms," *Medical image analysis*, vol. 31, pp. 63–76, 2016.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[11] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *arXiv preprint arXiv:1702.05747*, 2017.

[12] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *Jama*, vol. 316, no. 22, pp. 2402–2410, 2016.

[13] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[14] J. Oliveira and H. Proença, "Caries detection in panoramic dental x-ray images," in *Computational Vision and Medical Image Processing*. Springer, 2011, pp. 175–190.

[15] P. Lin, Y. Lai, and P. Huang, "An effective classification and numbering system for dental bitewing radiographs using teeth region and contour information," *Pattern Recognition*, vol. 43, no. 4, pp. 1380–1392, 2010.

[16] R. Ramteke and Y. K. Monali, "Automatic medical image classification and abnormality detection using k-nearest neighbour," *International Journal of Advanced Computer Research*, vol. 2, no. 4, pp. 190–196, 2012.

[17] A. Bosch, A. Zisserman, and X. Munoz, "Image classification using random forests and ferns," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[18] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," *Computer Vision–ECCV 2010*, pp. 143–156, 2010.

[19] H. Chen and A. K. Jain, "Tooth contour extraction for matching dental radiographs," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 3. IEEE, 2004, pp. 522–525.