Transfer learning-based super-resolution in panoramic models for predicting mandibular third molar extraction difficulty: a multi-center study

Wen Li1, Yang Li1, Xiao-Ling Liu1, Xiang-Long Zheng1, Shi-Yu Gao1, Hui-Min Huangfu1, Li-Song Lin1*

1Department of Oral and Maxillofacial Surgery, The First Affiliated Hospital of Fujian Medical University, Fuzhou 350005, China. 2School and Hospital of Stomatolgy, Fujian Medical University, Fuzhou 350002, China. 3Department of Oral and Maxillofacial Surgery, The First Affiliated Hospital of Zhengzhou University, Zhengzhou 450001, China. 4School of Mathematics and Statistics, Huazhong University of Science and Technology, Wuhan 430074, China.

Department of Dental Implantation, Suzhou Stomatological Hospital, Suzhou 215031, China.

*Corresponding to: Hui-Min Huangfu, Department of Dental Implantation, Suzhou Stomatological Hospital, No. 1366 Suzhan Road, Suzhou 215031, China. E-mail: hhfmi123456789@163.com. Li-Song Lin, Department of Oral and Maxillofacial Surgery, The First Affiliated Hospital of Fujian Medical University, No. 20 Cha Zhong Road, Fuzhou 350005, China. E-mail: dr_lsl@hotmail.com.

Author contributions

Wen Li and Yang Li contributed equally to this study. Wen Li and Yang Li wrote the main manuscript text. Xiaoling Liu collected information and data on the patients included in the study. Xianglong Zheng revised the manuscript text. Shiyu Gao analyzed data on the patients included in the study and revised the manuscript text. Humin Huangfu and Lisi Song Lin presented the research, oversaw its implementation. All authors reviewed the manuscript.

Competing interests

The authors declare no conflicts of interest.

Abstract

Background: This study aims to predict the extraction difficulty of mandibular third molars based on panoramic images using transfer learning while employing super-resolution (SR) technology to enhance the feasibility and validity of the prediction. Methods: We reviewed a total of 608 preoperative mandibular third molar panoramic radiographs from two medical facilities: the First Affiliated Hospital of Zhengzhou University (n = 509; 456 in the training set and 53 in the test set) and the Henan Provincial Dental Hospital (n = 99 in the validation set). We conducted a deep-transfer learning network on high-resolution (HR) panoramic radiographs to improve the longitudinal resolution of the images and obtained the SR images. Subsequently, we constructed models named Model-HR and Model-SR using high-dimensional quantitative features extracted through the Least Absolute Shrinkage and Selection Operator (LASSO) or transfer learning. The models' performances were evaluated using the receiver operating characteristic curve (ROC). To assess the reliability of the model, we compared the results from the test set with those of three dentists. Results: Model-SR outperformed Model-HR (area under the curve (AUC): 0.779, sensitivity: 85.5%, specificity: 60.9%, and accuracy: 79.8% vs. AUC: 0.753, sensitivity: 73.7%, specificity: 73.9%, and accuracy: 73.7%) in predicting the difficulty of extracting mandibular third molars. Both Model-HR (AUC = 0.821, 95% CI 0.687–0.956) and Model-SR (AUC = 0.963, 95% CI 0.921–0.999) demonstrated superior performance compared to expert dentists (highest AUC = 0.799, 95% CI 0.671–0.927). Conclusions: Model-SR yielded superior predictive performance in determining the difficulty of extracting mandibular third molars when compared with Model-HR and expert dentists' visual assessments.

Keywords: super-resolution, transfer-learning, mandibular third molar, extraction difficulty, panoramic radiographs
Introduction

The eruption of mandibular third molars is frequently impeded by adjacent teeth, bone tissue, or soft tissue obstructions [1]. Meanwhile, the primary cause of dental obstruction is the disparate degeneration of the jaw and the number of teeth as humans evolve, which results in a relatively smaller amount of bone than teeth and a lack of sufficient space in the jaw to accommodate all permanent teeth. Impacted teeth can lead to complications such as pericoronitis and infection in the maxillofacial space; thus, it is crucial to remove them at an early stage [2]. Extraction of mandibular third molars is a frequently performed operation in alveolar surgery, with a wide range of difficulty [3]. Due to the mandibular third molars' close proximity to the inferior alveolar nerve canal and the inherent anatomical variability of the nerve [4], precise clinical evaluation of extraction difficulty is critical for a successful operation. Conversely, the failure to accurately assess its complexity can result in inadequate surgical preparation and wasted time, as well as increased trauma and postoperative reactions [5, 6]. Various scoring systems have been developed to evaluate the complexity of mandibular third molar extractions. Nevertheless, there is no widely accepted method for assessing the difficulty of these extractions. To address this issue, we conducted an experiment introducing the modified Pederson index to categorize the difficulty of extracting patients’ mandibular third molars [7]. The Pederson index, based on Pell-Gregory classification [8], and Winter's scale serve [9] as initial methods for assessing extraction difficulty, but numerous researchers have proposed modified scales [10, 11]. Uniform criteria for predicting mandibular third molar extraction difficulty have proven elusive. In response, we employed a modified Pederson scale, dividing scores into two categories: ≤ 5 and > 5. The purpose of this method is to be more conducive to the application of clinical doctors. In the realm of Oral and Maxillofacial Surgery, several examination instruments are commonly employed, including magnetic resonance imaging (MRI), computed tomography (CT), panoramic radiography, and cone-beam computed tomography (CBCT). Among these modalities, panoramic radiography finds extensive clinical application, particularly in alveolar surgery [12]. Simultaneously, the field of artificial intelligence (AI) is undergoing rapid development, and one of its most promising branches, deep learning, is making significant progress in the medical domain [13–15]. Convolutional neural networks (CNNs), which are widely used in deep learning, have also found extensive use in stomatology. Currently, deep learning is utilized in the field of dentistry for diagnosing dental caries [16], periodontitis [17], cystic lesions [18], and tumors [19]. Despite these noteworthy advances, it is important to recognize that contemporary medical imaging radiomics signatures still grapple with challenges such as anisotropic resolution and suboptimal voxel statistics [20]. In an effort to augment the specificity and sensitivity of radiomics models, it becomes imperative to address the aforementioned challenges. One potential solution involves incorporating higher resolution images during the machine learning model construction process, which could contribute to more precise and accurate diagnostic outcomes.

Ongoing in the 1980s, super-resolution (SR) technology aims to recover higher spatial resolution in digital images from lower-resolution observations [21]. In recent years, the advent of deep learning (DL) has facilitated exceptional performance in medical imaging applications for SR [22, 23]. For instance, Niaz et al. [24] enhanced brain morphology diagnostic image quality through a DL-SR algorithm, which notably improved the visualization of fine structural details in elderly brains. The images recovered by SR not only demonstrate remarkable stability and reliability across multi-space ladder diagrams but also exhibit highly repeatable and robust radiomics features, as evidenced by several studies [20]. More recently, Mohammad et al. [25] discovered that employing SR images in oral panoramas can effectively improve image quality. However, to date, no studies have explored the application of DL-SR technology in predicting extraction difficulty for alveolar surgery. In the present study, our primary objective was to develop and validate a CNN-based model for the preoperative prediction of mandibular third molar extraction difficulty. Subsequently, we incorporated super-resolution reconstruction technology to enhance image quality. Furthermore, we endeavored to substantiate the model’s efficacy and evaluate its accuracy using data gathered from multiple centers. To guarantee a thorough assessment, we contrasted the model's performance with that of three dentists, each possessing varying qualifications, in determining tooth extraction difficulty.

Methods

Patients

Figure 1 illustrates the patient registration process. A total of 608 oral panoramic radiography images at the First Affiliated Hospital of Zhengzhou University and Henan Provincial Dental Hospital between May 2022 and October 2022 were eligible for inclusion. The inclusion criteria were as follows: (1) the presence of at least one mandibular third molar; (2) a clearly visible oral panoramic image; (3) age ≥ 16 years. The exclusion criteria were as follows: (1) the presence of significant metal artifacts in the images; (2) the involvement of cysts or tumors in the mandibular third molar; (3) missing data such as gender and age; (4) severely decayed mandibular third molars. The 456 mandibular third molars from the First Affiliated Hospital of Zhengzhou University were randomly selected as the training set and 53 were set aside for the test set. A total of 99 mandibular third molars from the Henan Provincial Dental Hospital were selected as the validation set.

Image acquisition and processing

The radiographic equipment used in this study included the Carestream 8000c (USA) at the First Affiliated Hospital of Zhengzhou University and the Carestream 9000c (USA) at the Henan Provincial Dental Hospital. The panoramic images were acquired with a voltage of 73 kV and a current of 6.3 mA. The acquired panoramic images were stored in portable network graphics (PNG) format in the image archiving and communication system. To process the 1500 × 750 pixels panoramic image into learnable data, pre-processing was required. The images were extracted from the panoramic image at a size of 200 × 200 pixels. Building upon the acquired panoramic image, we employ a deep transfer learning network to enhance the longitudinal resolution by quadrupling the pixel count (the pixels of 800 × 800). For detailed methods on SR technology, please refer to Supplementary Methods. First, Gaussian noise was added to the panoramic image to reduce the out-plane resolution with a factor of 4 to generate a new low-resolution image. Then, the low-resolution and synthetic high-resolution (HR) image pairs were used to train a lightweight parallel generative adversarial network (GAN) model. Finally, the trained model was applied to HR by transfer learning. Consequently, the resulting images are referred to as SR images [26].

Study design

In this study, the model development process is mainly divided into four stages: source data pre-training, transfer learning, target data fine-tuning training, and model performance verification. The model process is depicted in Figure 2. Initially, the source data is pre-trained

Figure 1 Flow diagram of the study population
Figure 2  Workflow of the study. HR, high-resolution. SR, super-resolution.

<table>
<thead>
<tr>
<th>Table 1  Pederson scale</th>
<th>Type</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angulation</td>
<td>Mesioangular</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Horizontal/Transverse</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Vertical</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Distoangular</td>
<td>4</td>
</tr>
<tr>
<td>Depth</td>
<td>Level A: high occlusal level</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Level B: medium occlusal level</td>
<td>2</td>
</tr>
<tr>
<td>Ramus</td>
<td>Level C: deep occlusal level</td>
<td>3</td>
</tr>
<tr>
<td>relationship</td>
<td>Class1: sufficient space</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Class2: reduced space</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Class3: no space</td>
<td>3</td>
</tr>
</tbody>
</table>

to obtain the weight parameters of the model. Subsequently, the remaining network layers of the new model are initialized with the obtained model weight parameters. Following that, the target dataset is fine-tuned to acquire the final model, and ultimately, the model’s performance is tested on the validation set.

DL algorithms were constructed using Python. This research employs the ResNet152 structure through transfer learning [27]. The network inputs are the pixels of 200 × 200 and 800 × 800. During the experiment, the model is pre-trained on the ImageNet dataset to obtain a model.

The model was subsequently assessed visually to comprehend its functioning. The classification activation heatmap was employed to emphasize the most crucial anatomical regions utilized by the model when categorizing images. The classification activation heatmap is a thermal gradient map, where the warmer colors signify the most significant regions in the classification process.

The Pederson Scale in Table 1 was utilized in combination with the difficulty scores, where scores ≤ 5 were grouped as one category, and > 5 were grouped as another.

Feature extraction

The model training involved updating network weights through the utilization of a cross-entropy loss function to address the prediction task. To optimize this process, an adaptive moment estimation optimizer with a learning rate of 0.1 was employed for 30 epochs, using a batch size of 64. Once the training was concluded, the network parameters were fixed, and the resulting model was utilized as a feature extractor. Specifically, 2048 DL features were extracted from the penultimate layer of the fine-tuned ResNet152 for each patient within both the training and validation cohorts, as detailed in the Supplementary Datasheet 1. Subsequently, a visual assessment was conducted to comprehend the operational characteristics of the model.

Feature selection

We also conducted statistical tests using the Mann-Whitney U test and performed feature screening on all features. Only features with a p-value of less than 0.05 were retained for further analysis.

Features with high repeatability were subjected to Spearman’s rank correlation coefficient to calculate their correlation. Only one feature with a correlation coefficient greater than 0.9 between any two features was retained to ensure the diversity of the feature set [28]. To retain the most informative features, a greedy recursive deletion strategy was used for feature filtering. Specifically, the feature with the largest redundancy in the current set was deleted one at a time. After the filtering process, 32 features were retained for subsequent analysis.

In this study, we used the Least Absolute Shrinkage and Selection Operator (LASSO) regression model for the feature-constructed discovery dataset. LASSO shrinks all regression coefficients to zero, depending on the adjustment weight λ, and sets the coefficients of many uncorrelated features to exactly zero. We used 10-fold cross-validation and a minimum criterion to find the optimal λ, where the final value of λ produced the minimum cross-validation error. We only retained features with non-zero coefficients for regression model fitting and combined them to form features.

After LASSO feature screening, the final features were input into various machine learning models, including Logistic Regression (LR), NaiveBayes, K-Nearest Neighbor (KNN), AdaBoost, and Multilayer Perceptron (MLP), for constructing the model.

Comparison with dentists

To evaluate the performance of the model, three dentists were recruited to rate the test set according to the Pederson Scale. Prior to diagnosing the test data, the dentists underwent a training session, which involved an explanation of the scoring scale and a demonstration of six demo cases. The test set consisted of 53 cases that were not included in the dataset.

Statistical analysis

The predictive performance of the models was evaluated through the utilization of receiver operating characteristic (ROC) curves. The area under the ROC curve (AUC) was computed to quantify the discriminative capability, while the balanced sensitivity and specificity were determined based on the cut-off point yielding the maximum Youden index value. To establish the 95% confidence interval (CI) for the AUC, the bootstrap method was employed with 1000 intervals. The AUC, ranging from 0.5 to 1.0, serves as a metric for assessing the discriminatory power of the test. A value of 1.0 signifies a perfect discriminant test, whereas a range between 0.8 and 1.0 indicates a good discriminant test. A moderate discriminant test is identified when the AUC falls between 0.6 and 0.8, whereas a poor
Figure 3 Receiver operating characteristic (ROC) curves of the different model convolutional neural network (CNN) data set. A: high-resolution group. B: super-resolution group.

Figure 4 Heat map with the AUC of different models. HR, high-resolution; SR, super-resolution; LR, Logistic Regression; KNN, K-Nearest Neighbor; MLP, Multilayer Perceptron; AUC, area under curve.

discriminant test is characterized by an AUC ranging from 0.5 to 0.6 [29]. Statistical analyses were conducted using SPSS software (version 21.0), with statistical significance defined as a two-sided p-value ≤ 0.05.

To evaluate the performance of the classification model, we calculated sensitivity, specificity, and accuracy. These metrics are calculated based on the concepts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

\[
\text{Sensitivity} = \frac{TP}{TP + FN} \\
\text{Specificity} = \frac{TN}{TN + FP} \\
\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}
\]

Ethics approval and consent to participate

The criteria and procedures for the inclusion of participants in this study were in accordance with the criteria indicated in the 1964 Helsinki Declaration. This study was approved by the Ethics Committee of The First Affiliated Hospital of Zhengzhou University (approval number: 2022-KY-0285). The study was retrospective and therefore did not require written informed consent from the participants. The informed consent waiver for this study was approved by the Ethics Committee of The First Affiliated Hospital of Zhengzhou University. And the informed consent waiver does not infringe on the rights and life and health of the participants.

Results

Patient characteristics

A total of 317 patients with a mean age of 37.85 ± 14.82 (age range of 16–79 years, 118 male, 199 female) with a total of 509 mandibular third molars were found in the First Affiliated Hospital of Zhengzhou University. The Henan Provincial Dental Hospital had a total of 63 patients with a mean age of 30.84 ± 9.31 (age range of 18–58 years, 19 male, 44 female) with 99 mandibular third molars included in the validation set.

Predictive performance of the models

Figure 3 presents the results of five distinct machine learning algorithms applied to both Model-HR and Model-SR. In the case of Model-HR, the most effective machine learning algorithm was MLP, which yielded an area under the curve (AUC) of 0.753 (95% CI 0.640–0.866); further details can be found in Supplementary Table 1. Besides, for Model-SR, the optimal machine learning algorithm was LR, achieving an AUC of 0.779 (95% CI 0.660–0.899) with additional information provided in Supplementary Table 2. A heatmap format in Figure 4 displays the AUC across these models. Notably, the AUC results for the Model-SR group validation set, utilizing the SR reconstruction technique, surpass those of the Model-HR group. Both Model-HR (AUC = 0.821, 95% CI 0.687–0.956) and Model-SR (AUC = 0.963, 95% CI 0.921–0.999) demonstrated superior performance compared to expert dentists (highest AUC = 0.799, 95% CI 0.671–0.927). A comprehensive evaluation of accuracy, sensitivity, and specificity is outlined in Table 2.

Classification activation heatmap

Figure 5 illustrates the class activation maps for Model-HR and Model-SR classifications. In these maps, the red color represents the most influential area, while the blue color denotes the least influential area in the decision-making process. Examining the model established by Model-HR (Figure 5A), it becomes evident that, apart from the tooth's crown appearing blue (not recognized), the remaining area is red. Furthermore, the model established by Model-SR (Figure 5B) displays a blue area that clearly covers the entire tooth, with the region around the crown being entirely encompassed by red. This observation indicates that, after employing the SR reconstruction technique, the model is better equipped to identify and interpret key parts in the image, consequently yielding more accurate results.

Discussion

In recent years, deep learning models have been proposed and
demonstrated impressive results across various domains, sometimes even surpassing clinicians in prediction accuracy [30]. Although deep learning’s application in dentistry is relatively new, there is growing interest in leveraging this technology for dental imaging. Current diagnostic tests for detecting oral diseases include periapical radiographs [31], panoramic radiographs [32], and CBCT [33]. For instance, Lee et al. [31] employed a pre-trained GoogleNet Inception v3 network model to predict dental caries using a dataset of 3,000 periapical radiographs, achieving high accuracy. Similarly, Lee et al. [34] study reveals that deep CNNs exhibit superior diagnostic accuracy in identifying periodontally compromised teeth in premolars compared to molars. Furthermore, when combined with a clinical model, the deep learning model generated predictions for the 5-year risk of implant loss in patients that were on par with skilled clinicians [35]. The AUC for this study was reported to be 0.90 (95% CI, 0.84–0.95). These findings hold significant value for implantologists, as they offer a reliable foundation for preoperative assessment of surgical risk and prognosis. By harnessing deep learning in dental imaging, we can improve diagnostic accuracy and provide better patient care.

In a study related to predicting the difficulty of mandibular third molar extraction and associated complications, Vinahalingam et al. [36] employed the U-net method for automating the segmentation of mandibular third molars and nerve canals. This approach facilitated clinical decision-making in assessing the risk of inferior alveolar nerve injury following third molar extraction. In another study, Yoo et al. [37] collected 1,053 mandibular third molars and utilized a CNN deep learning model to predict the extraction difficulty of mandibular third molars. They achieved accuracies of 78.91%, 82.03%, and 90.23% in depth, ramal relationship, and angulation, respectively. Furthermore, Lee et al. [38] combined the two methods and not only predicted extraction difficulty but also assessed the risk of inferior alveolar nerve injury using an automated detection approach. Unlike previous studies in this area, which did not incorporate SR techniques, our study developed and independently validated a HR and SR-based model for predicting mandibular third molar extraction difficulty. Model-SR demonstrated superior predictive performance and enhanced image quality compared to clinicians and Model-HR, suggesting its potential to assist clinicians in choosing the best initial treatment strategy for patients. Currently, the lower resolution of video images may have adverse effects, which can be addressed by super-resolution technology [39, 40]. This method has been proposed for numerous medical imaging applications and can mitigate issues such as high-frequency information loss, edge blurring, and improved resolution [41, 42]. In our aforementioned study, we developed a GAN-based SR model and introduced its application in predicting the difficulty of mandibular third molar extraction in alveolar surgery. Our model exhibited excellent diagnostic performance (AUC = 0.779, 95% CI: 0.660–0.899). Figure 5 demonstrates a noticeable improvement in edge clarity within the SR-processed image. This enhanced clarity suggests that DL-based image recognition yields more accurate results. Notably, the blue area distinctly encompasses

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Accuracy (%)</th>
<th>AUC</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model HR-MLP</td>
<td>0.950</td>
<td>0.615</td>
<td>0.868</td>
<td>0.821</td>
<td>0.687-0.956</td>
</tr>
<tr>
<td>Model SR-LR</td>
<td>0.850</td>
<td>1.000</td>
<td>0.887</td>
<td>0.963</td>
<td>0.921-0.999</td>
</tr>
<tr>
<td>Dentist I</td>
<td>0.500</td>
<td>0.923</td>
<td>0.604</td>
<td>0.712</td>
<td>0.565-0.858</td>
</tr>
<tr>
<td>Dentist II</td>
<td>0.600</td>
<td>0.847</td>
<td>0.661</td>
<td>0.723</td>
<td>0.572-0.875</td>
</tr>
<tr>
<td>Dentist III</td>
<td>0.675</td>
<td>0.923</td>
<td>0.736</td>
<td>0.799</td>
<td>0.671-0.927</td>
</tr>
</tbody>
</table>

HR, high-resolution; SR, super-resolution; MLP, Multilayer Perceptron; LR, Logistic Regression; AUC, area under receiver operating characteristic curve; CI, confidence interval.

Figure 5 Classification activation heatmap. A: high-resolution group. B: super-resolution group.
the entire tooth, while the red area effectively covers the surrounding crown region. This observation underscores the effectiveness of the SR reconstruction technique in facilitating the model's enhanced identification and interpretation of crucial image components, leading to improved accuracy in the obtained results. SR techniques, transcending the physical limitations of imaging systems, offer two key advantages: Firstly, they improve overall image quality by reducing noise, artifacts, and other distortions present in the original image. Secondly, increasing image resolution reveals finer details that might have been missed in the original image, making it easier to identify specific features or patterns.

In contrast to single-center studies, our multi-center experiment utilized various equipment, enhancing the model's generalizability and robustness. The results indicate that the proposed deep learning-based model can effectively predict mandibular third molar extraction difficulty. In fact, the model outperformed dentists in diagnosing the test set. While the models demonstrated high sensitivity and low false negative rates, physicians exhibited high specificity and low false positive rates. This discrepancy may stem from surgeons’ subjective input during the judgment process, leading to misclassification of some impacted teeth that appear complex on panoramic film but prove simpler during surgery. Moving forward, deep learning analysis of digital panoramic images could serve as an adjunct diagnostic tool for assessing mandibular third molar extraction difficulty.

Although the model achieved a high accuracy rate, our study has some limitations that need to be addressed. Firstly, a major limitation is the small sample size that we have collected, which can affect the accuracy of the model built using DL. Secondly, the current diagnostic model does not incorporate important clinical information such as age, gender, number of roots, and root morphology, which also play a crucial role in assessing the extraction difficulty of the mandibular third molar. In conclusion, we recognize that the cases included in our study were largely limited to individuals in China. As a result, we were unable to determine if there are disparities in the difficulty of mandibular third molar extraction across different ethnic groups. Future studies that span across diverse populations could help address this important gap in knowledge. It is important to include these factors in future studies to improve the accuracy and generalizability of the model.

Conclusion

In conclusion, the study presented a DL-based SR model, which outperformed the conventional model and expert dentists in predicting the difficulty of extracting mandibular third molars. So far, this is the first time the DL-based SR method has been applied to analysis for dentist decision support. Incorporating artificial intelligence and machine learning in medical research will assist clinicians in solving problems in the future.

References


