Zhu et al. Journal of Translational Medicine (2024) 22:930
https://doi.org/10.1186/s12967-024-05741-3

Journal of Translational Medicine

RESEARCH Open Access

LMCD-OR: a large-scale, multilevel categorized diagnostic dataset for oral radiography

Jiaqian Zhu[1,2†], Li Zeng[3†], Zefei Mo[4†], Luhuan Cao[5†], Yanchan Wu[6], Liang Hong[7*], Qi Zhao[8,9*] and Feifei Su[1,10,11*]

Abstract
In recent years, digital dentistry has increasingly utilized advanced image analysis techniques, such as image classification and disease diagnosis, to improve clinical outcomes. Despite these advances, the lack of comprehensive benchmark datasets is a significant barrier. To address this gap, our research team develop LMCD-OR, a substantial collection of oral radiograph images designed to support extensive artificial intelligence (AI)-driven diagnostics. LMCD-OR comprises 3,818 digital imaging and communications in medicine (DICOM) oral X-ray images from local medical institutions that are meticulously annotated to provide broad category information for both primary dental outpatient services and detailed secondary disease diagnoses. This dataset is engineered to train and validate multiclassification models to improve the precision and scope of oral disease diagnostics. To ensure robust dataset validation, we employ four cutting-edge visual neural network classification models as benchmarks. These models are tested against rigorous performance metrics, demonstrating the ability of the dataset to support advanced image classification and disease diagnosis tasks. LMCD-OR is publicly available at http://dentaldataset.zeroacademy.net.

Keywords Dentistry, Dataset competition, AI-driven diagnosis, Multilevel classification, Baseline models

†Jiaqian Zhu, Li Zeng, Zefei Mo and Luhuan Cao contributed equally to this work and share first authorship.

*Correspondence:
Liang Hong
rahoshongliang@163.com
Qi Zhao
zhaoqi@lnu.edu.cn
Feifei Su
feifeisuzs@163.com
[1]Department of Infectious Diseases, Wenzhou Central Hospital, Wenzhou 325000, China
[2]The First School of Medicine, School of Information and Engineering, Wenzhou Medical University, Wenzhou 325001, China
[3]The Second Clinical Medical College of Wenzhou Medical University, Wenzhou 325000, China
[4]School of Biomedical Engineering, School of Ophthalmology and Optometry, Eye Hospital, Wenzhou Medical University, Wenzhou 325000, China
[5]School of Nursing, Wenzhou Medical University, Wenzhou 325001, China
[6]School of Electrical and Information Engineering, Quzhou University, Quzhou 324000, China
[7]Department of Infectious Diseases, The Third Affiliated Hospital of Wenzhou Medical University, Wenzhou 325000, China
[8]School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China
[9]Wenzhou Institute, University of Chinese Academy of Sciences, Wenzhou 325000, China
[10]Department of Infectious Diseases, Wenzhou Sixth People's Hospital, Wenzhou 325000, China
[11]Wenzhou Key Laboratory of Diagnosis and Treatment of Emerging and Recurrent Infectious Diseases, Wenzhou 325000, China

Introduction

Oral health is crucial for maintaining overall well-being [1]. However, oral diseases continue to pose significant global health challenges [2]. According to the 2019 Global Burden of Disease (GBD) report, oral diseases such as dental caries, periodontitis, and tooth loss affect more than 44.5% of the global population, making them a major public health burden and a key determinant of quality of life [3]. Furthermore, a 2021 review on global oral health disparities highlights that the burden of oral diseases is particularly severe in low- and middle-income countries due to a lack of access to dental care services [4]. There is a significant gap between the actual burden of disease and the healthcare system's ability to provide adequate patient care. These findings underscore the urgent need to enhance the effectiveness of oral disease diagnosis, improve dental services, and conduct data-driven research to mitigate the growing global impact of oral diseases. Therefore, there is an urgent need for a comprehensive oral health dataset that includes a large volume of clinical data. Such a dataset would not only provide more accurate guidance for diagnosing and treating oral diseases in specific regions but also serve as a foundation for developing artificial intelligence (AI) models. In recent years, AI algorithms have experienced rapid and significant advancements [5–9]. By leveraging these datasets, AI algorithms can more effectively detect patterns and abnormalities in oral health, thus providing fast and accurate diagnostic support. The integration of AI with well-curated datasets holds great potential to significantly reduce diagnostic and treatment time, ultimately improving clinical outcomes [10, 11].

Oral datasets have shown considerable potential in dental informatics [12], facilitating early detection and diagnosis of oral diseases and assisting in the identification of osteoporosis and anatomical landmarks within gingival and periodontal tissues [13]. For example, datasets comprising oral X-ray images can be utilized to train AI models to detect early signs of dental caries and periodontal diseases more accurately. Such models have been demonstrated to assist clinicians in identifying subtle radiographic features that may be overlooked during manual examination, thereby improving diagnostic accuracy and efficiency [14]. Overall, integrating these datasets into clinical workflows not only enhances diagnostic precision but also contributes to more timely and personalized treatment planning, representing a significant advance in intelligent oral healthcare. The development and exploration of these datasets represent important focuses for advancing future precision-oriented dental healthcare to significantly improve the efficiency and accuracy of dental practices. To meet this objective, various datasets, such as the Tufts Dental Database (TDD) [12], the oral diseases and sciences image database (ODSI-DB) [13], the Swedish Quality Registry for Caries and Periodontal Diseases (SKaPa) dataset, the National Dental Practice-based Research Network (NDPBRN) dataset, and the Bigmouth dataset [15], have been developed to assist in diagnosing various dental diseases. The existing oral datasets fall into three primary categories: oral cancer databases, which are represented by organizations such as the Queensland Cancer Registry (QCR) and can predict the 3-year and 5-year overall survival of oral cancer patients [16]; oral radiographic datasets, which are represented by organizations such as authority (AUTH) and are proficient in accurately dissecting intraoral anatomical structures [17]; and intraoral photographic datasets, which are represented by organizations such as the University of Hong Kong/Hospital Authority Hong Kong West Cluster (HKU/HA HKW IRB) and are adept at categorizing, labeling and locating gingival health status [18].

However, the existing oral datasets are relatively limited in scope and size, and a sufficiently large dataset that can serve as a baseline for comparing different models is lacking [19]. Most studies have generally relied on single models, such as the standard convolutional neural network (CNN) model and the faster region-based convolutional neural network (R-CNN), for training, which results in unstable raw data processing [20]. Within the current oral radiograph dataset, Panetta created the TDD, which contains 120 periapical dental radiographs from both the maxilla and mandible. This database utilized CNNs to analyze bone conditions, achieving a nuanced assessment of baseline results in dental radiographic image enhancement and tooth segmentation. However, the limited size of the database, its reliance on a single model, and its relatively poor generalizability prevent its clinical translation [12]. While Mattea et al. expanded the database to include 2112 radiographs from seven key institutions in an attempt to reduce bias and the misinterpretation of results, relying solely on CNNs for data analysis remains a challenge in ensuring model efficacy and stability [21]. Mohammed et al. applied six pretrained models, AlexNet, VGG-16, VGG-19, ResNet-50, DenseNet-169 and MobileNet-V3, to 500 cone beam computed tomography (CBCT) images. Although successful in accurately detecting the location of missing teeth, these models cannot simultaneously detect other oral diseases [12]. Many oral datasets lack stand-alone websites, limiting the expansion of competition and research communities [22–24]. While few databases, such as the TDD, have easily accessible stand-alone websites, they primarily process disease types via single model identification. Although models have been developed to identify multiple diseases over the past three years, their ability to detect a wide array of diseases is still somewhat limited and often restricted to recognizing only approximately 20 different categories [25, 26].

Current models for oral databases face several challenges: they are based on relatively small datasets, lack dedicated website support, and often rely on training models that are too uniform. These factors obstruct the creation of oral diagnostic tools with broad applicability. To overcome these obstacles, increasing the data volume significantly, setting up specialized websites for easy access, and encouraging the adoption of various foundational models for the processing of data are imperative. These steps improve model generalizability and promote the sharing of resources.

LMCD-OR addresses the challenges faced by current oral databases in several ways. First, it expands the data volume by incorporating 3,818 oral radiograph images sourced from local hospital databases. These images include original DICOM oral X-rays that have been meticulously labeled to cover both common categories seen in primary dental care and specific secondary disease diagnoses, thereby enhancing the dataset's representativeness and reliability for various diagnostic tasks. Second, to tackle the issue of accessibility, LMCD-OR is made available through dedicated websites such as http://dentaldataset.zeroacademy.net/ [27] and Kaggle [28]. These platforms provide convenient access for researchers and clinicians to explore, download, and utilize the dataset, fostering community collaboration and resource sharing. To ensure transparency, we emphasize that the dataset is provided under open access, with clear usage terms and conditions outlined on each platform. Users must agree to non-commercial use only, and they are required to acknowledge the dataset in their publications. Additionally, the legal responsibilities for appropriate use are clearly stated, ensuring that users are aware of their obligations when accessing the dataset. Finally, to address the challenge of overly uniform training models, we have developed a baseline multiclass classification model using diverse foundational architectures. This model not only aids in evaluating the dataset's performance but also sets a benchmark for future research and competitions. These comprehensive efforts ensure that LMCD-OR serves as a robust and versatile resource for advancing oral health diagnostics and treatment outcomes [29].

Furthermore, the aim of compiling oral datasets on a specialized webpage and community, as part of this study, is to invigorate the development of competitions based on oral disease databases while nurturing a research community focused on this field.

Method

Ethics and quality standards

The subjects included in this study received approval from the Medical Ethics Review Committee of the Affiliated Stomatological Hospital of Wenzhou Medical University (Ethics No. WYKQ2023010). Owing to the retrospective nature of this investigation and the anonymity and nonidentifiability of the radiographs, the institutional review board granted a waiver of informed consent. In addition, our study methodology and reporting strictly adhere to the standards for quality improvement reporting excellence (SQUIRE) guidelines, ensuring ethical and quality standards for this quality improvement study.

Datasets

Oral radiographic data and case data were collected between 1st July 2023 and 18th August 2023, comprising a total of 4682 raw DCM files and 79,139 raw case data. The case data collection primarily includes 288 principal diagnoses of diseases and seven medical operations, with detailed tables of these diagnoses presented in Table S1 in the Supplementary Materials. All the raw images are stored at a uniform resolution (2440×1280). The DCM format is converted to a portable network graphics (PNG) format, incorporating several levels of data enhancement for subsequent data processing. This includes measures such as contrast enhancement, image flipping and image rotation.

Inclusion and exclusion criteria

Images are rigorously screened, and DICOM images without corresponding condition information or with multiple mixed information segments are systematically excluded. Images of elderly patients over 75 years of age and children under 4 years of age are also excluded. All images with a primary diagnosis of medical surgery are removed, along with case records with a primary diagnosis associated with fewer than five patients. Our dataset contains 3818 unique oral radiographs and an equal number of individual heat codes. The flow of images and annotations into and out of this set is shown in Fig. 1.

Image annotation

Primary dataset annotation was undertaken from 24th November 2023 to 8th December 2023, followed by a secondary review and quality control phase from 9th January 2024 to 29th January 2024. Annotation at the image level was performed by four certified dental clinicians. Subsequent quality control of the entire annotated sample was completed by three quality controllers. Prior to the start of the annotation process, a group of medical experts highlighted seven initial superordinate detection targets and 41 initial subordinate detection targets. These categories are detailed in Table S2.

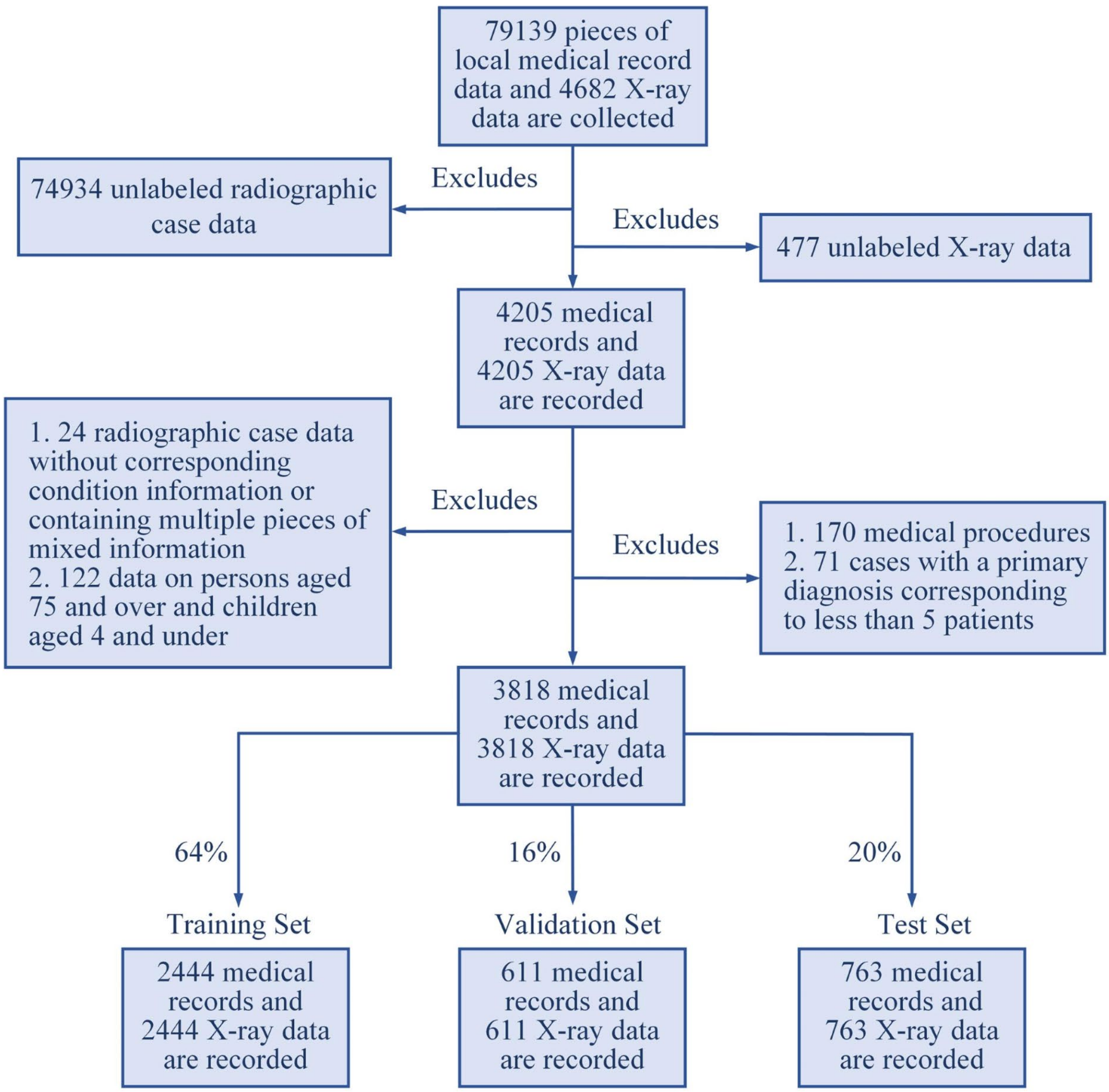

Fig. 1 Data inclusion and elimination flowchart

Training set, validation set and test set

The dataset is methodically divided into three distinct segments, ensuring balanced proportions and no patient overlap between the subsets: the training set comprises 64% with 2444 images, the validation set comprises 16% with 611 images, and the test set comprises 20% with a total of 763 images. The test set is stored in an isolated storage space managed by professional data scientists (HM. C and FX.C). In the competition setting, this test set is exclusively used to evaluate algorithm performance when a user submits their algorithm for the competition. The descriptive statistics for all the subsets are shown in Table 1. The dataset is publicly available on a dedicated website [27] as well as on Kaggle [28].

Benchmarking algorithm

Algorithm design

In this project, we create a sophisticated multiclassification model designed for automatically analyzing oral X-ray images to demonstrate the efficacy of LMCD-OR in both image classification and disease diagnosis tasks. To achieve this goal, we employ a transfer learning methodology using deep learning models pretrained on extensive image recognition datasets (such as ImageNet).

Table 1 Demographic table

Clinical parameters	Data	Training set ($n=2444$)	Validation set ($n=611$)	*p* value
Age (years)	27.05 ± 17.98	27.22 ± 18.05	26.33 ± 17.72	0.381 > 0.05
Gender				
(male)	1339(0.37)	1053(0.37)	286(0.39)	0.256 > 0.05
(female)	1715(0.63)	1390(0.63)	325(0.61)	
Type of registration				
(first visit)	2249(0.63)	1798(0.63)	451(0.61)	0.992 > 0.05
(return visit)	806(0.37)	646(0.37)	160(0.39)	
Check-in time				
7:00——9:30	1318(0.43)	1068(0.44)	250(0.41)	0.929 > 0.05
9:31——12:00	407(0.13)	321(0.13)	85(0.14)	
12:01——14:30	946(0.30)	746(0.30)	200(0.33)	
14:31——17:00	385(0.13)	309(0.13)	76(0.12)	
Number of patients in each department				
Oral surgery disease	615(0.20)	492(0.20)	123(0.20)	0.999 > 0.05
Pediatric stomatology diseases	74(0.02)	58(0.02)	16(0.02)	
Remediation disease	286(0.09)	226(0.09)	60(0.09)	
Dental endodontic diseases	1003(0.33)	801(0.33)	202(0.33)	
Periodontal diseases	346(0.11)	276(0.11)	70(0.11)	
Orthodontic disease	722(0.24)	583(0.24)	139(0.25)	
Other	9(0.01)	8(0.01)	1(0.00)	
Types of dental diseases in each department				
Oral surgery disease	8(0.20)	8(0.20)	7(0.18)	0.999 > 0.05
Pediatric stomatology diseases	2(0.05)	2(0.05)	2(0.05)	
Remediation disease	3(0.07)	3(0.07)	2(0.05)	
Dental endodontic diseases	17(0.42)	17(0.42)	17(0.45)	
Periodontal diseases	3(0.07)	3(0.07)	3(0.08)	
Orthodontic disease	6(0.14)	6(0.14)	6(0.16)	
Other	2(0.05)	2(0.05)	1(0.03)	

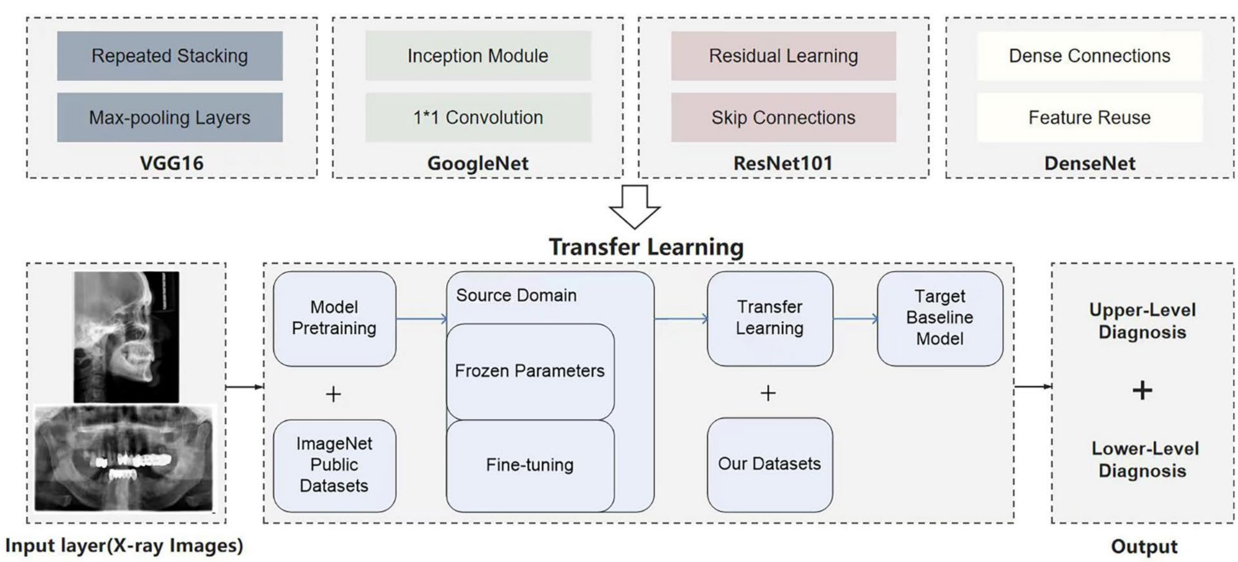

Fig. 2 Schematic representation of a transfer learning-based multiclass classification model

These models include VGG-16 [30], GoogLeNet [31], ResNet-152 [32], and DenseNet-201 [33], as depicted in Fig. 2. Each model is characterized by unique architectural features. VGG-16 is unique because it uses a consistently small convolution kernel. GoogLeNet achieves parameter minimization through its innovative “sparse connection” design. ResNet-152 addresses vanishing gradients in deep networks via deep residual learning. DenseNet-201 enhances feature propagation and reuses its dense connectivity architecture.

These pretrained models serve as initialization parameters and are fine-tuned specifically for our multiclassification task on oral radiograph data. The output layer includes a softmax classifier combined with a cross-entropy loss function to optimize the classification performance. The selection of the optimal model, which serves as the baseline model for both the superordinate diagnostic task and subordinate diagnostic task, is influenced by performance evaluation metrics such as the top-1 accuracy (top-1 Acc) and top-5 accuracy (top-5 Acc).

Experiment

Various data enhancement techniques are used, including image contrast enhancement, flipping, and rotation. The computational framework uses the stochastic gradient descent optimizer (SGD) with a maximum learning rate of 0.0001. The impulse technique is used with its value set at 0.9 and a weight decay parameter set at 0.05 to effectively manage the optimization process and accelerate the model convergence rate. The number of iterations for the weight adjustment process is limited to 150 epochs, and the model accuracy performance improvement converges when no change greater than 0.02 is observed for 15 consecutive epochs on the validation set. The final model is evaluated via an independent test set.

Statistical analysis

The top-1 Acc of the multiclassification task is calculated to evaluate the final performance of the baseline algorithm. The top-5 Acc is also calculated for the complex subordinate target classification task. In addition to these two metrics, we introduce six other metrics—microaveraged precision, microaveraged recall, microaveraged F1, macroaveraged precision, macroaveraged recall, and macroaveraged F1—to provide a more comprehensive performance assessment. The Mann‒Whitney U test is employed to evaluate the differences in performance between the groups. These comprehensive assessment metrics are analyzed via an open-source statistical tool based on Python. The specific versions of the Python library used can be found in Table S2.

Data desensitization

To ensure privacy, we implement a multistage desensitization process before the data are stored and distributed. First, numeric data are rounded and quantized to reduce their precision and sensitivity. Second, masking techniques are used to hide sensitive information and truncate data to reduce precision. In addition, a unique substitution method replaces the original data with irrelevant unique identifiers to prevent reverse inference. We also use hash functions to convert data into fixed-length strings to increase data security. To further break the correlation between data, a reordering process randomizes the order of records. Finally, format-preserving encryption (FPE) ensures that the data content is unrecognizable while preserving the data format. The combined application of these desensitization methods effectively reduces the risk of personal identification and microdata exposure while maximizing the retention of the research value and utility of the data.

Results

Data analysis of the local dataset

From 1 July 2023 to 18 August 2023, a total of 3,818 patients diagnosed with dental diseases were evaluated at a medical center in China. To improve model performance and ensure unbiased validation, thereby increasing the credibility of the study conclusions, patient selection and distribution were randomized across cohorts. The data distribution heatmaps for the six datasets are shown in Figure S1. Figure S1a displays a heatmap of metric attributes for 3,818 patients, covering medical status, registration time, sex, registration type, and main diagnostic categories. Medical status ranges from 'number returned' (0) to 'recalled' (6). Registration time is segmented into four periods: 7:00–10:00 (1), 10:01–13:00 (2), 13:01–16:00 (3), and 16:01–19:00 (4). Gender, registration type, and electronic signature are binary variables denoted 0/1. A female is assigned 0, and a male is assigned 1. The initial visit is denoted 0, and the returned visit is denoted 1. Unsigned is represented by 0, and signed is represented by 1. The main diagnostic categories range from 'oral surgery diseases' (0) to 'other' (6). Figures S1b, S1c, and S1d show heatmaps for metric attributes within the model development, validation, and external cohorts that were utilized in model optimization, respectively. Table 1 provides the data and between-group statistical test results. The model development cohort (n=2444), the validation cohort (n=611) and the external cohort used for model optimization (n=763) are shown in Figure S1b to S1d, respectively.

Model accuracy performance for superordinate classification tasks

The transfer learning models are constructed via the four replacement core frameworks. Their performances within the superordinate classification tasks are analyzed using both the validation and test sets, as shown in Fig. 3. In the validation set, the ResNet-152, VGG-16, DenseNet-201, and GoogLeNet frameworks provide top-1 Acc values of 0.6674, 0.6636, 0.6487, and 0.6404, respectively, as shown in Fig. 3a. We also calculated the microaveraged precision, microaveraged recall, microaveraged F1, macroaveraged precision, macroaveraged recall, and

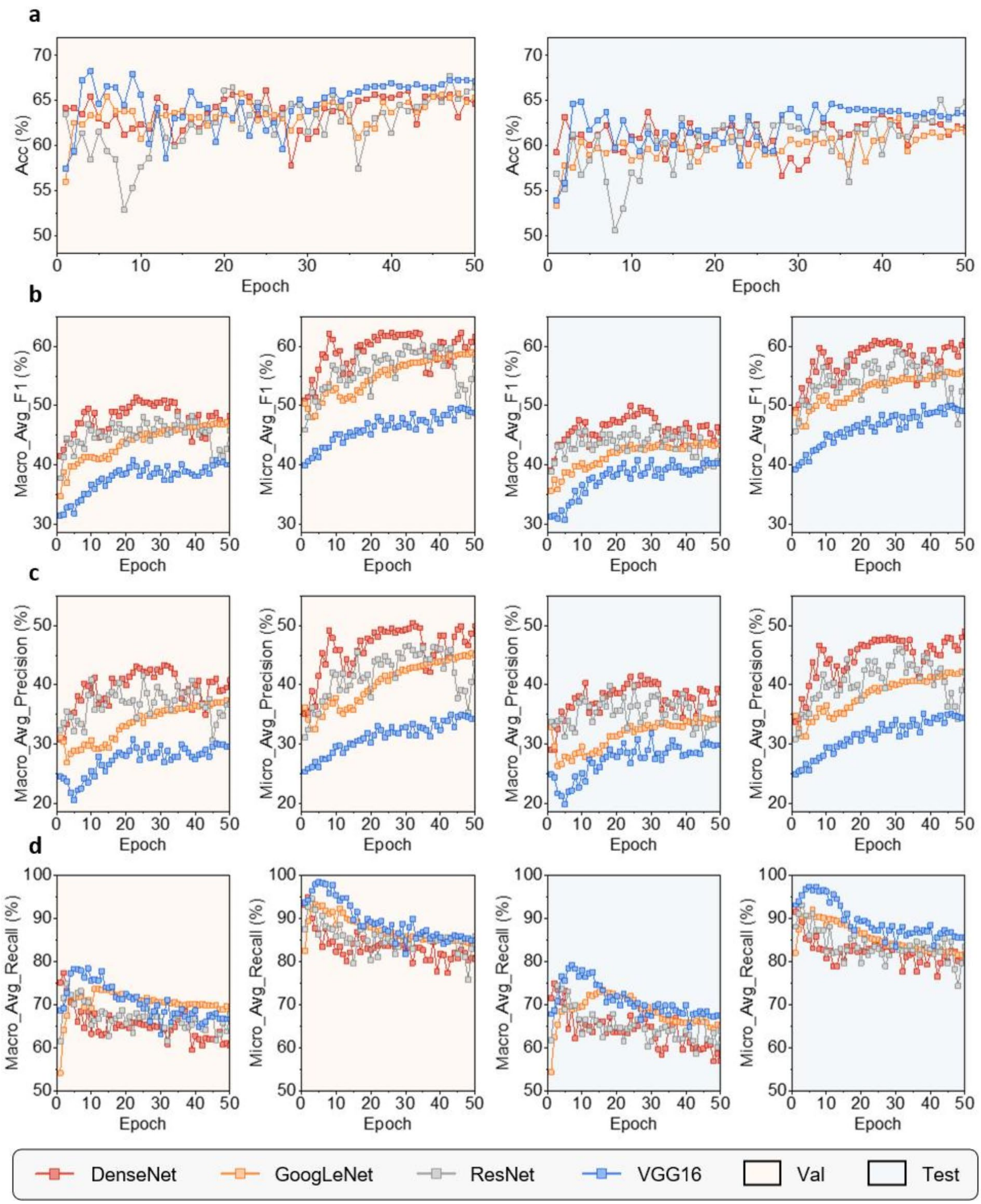

Fig. 3 The four transfer learning models using different core frameworks were evaluated on the validation and test sets via the metrics top-1 Acc (**a**), macroaveraged F1 and microaveraged F1 (**b**), macroaveraged precision and microaveraged precision (**c**), macroaveraged recall and microaveraged recall (**d**) for superordinate classification tasks

macroaveraged F1 within the ResNet-152 framework. The values are 0.4625, 0.8417, 0.5966, 0.3680, 0.6397, and 0.4577. See Fig. 3b and d. Conversely, in the test set, all the models represented by the four frameworks show a decrease in performance, with top-1 Acc values of 0.6445, 0.6358, 0.6236, and 0.6042 for the ResNet-152, VGG-16, DenseNet-201, and GoogLeNet frameworks, respectively, as shown in Fig. 3a. The DenseNet-201 framework shows the highest performance across several metrics, achieving a microaveraged precision of 0.5010, a microaveraged F1 of 0.6197, a macroaveraged precision of 0.4029, and a macroaveraged F1 of 0.4680. Conversely, the VGG-16 framework outperforms the other methods in terms of the microaveraged recall and macroaveraged recall, with scores of 0.8575 and 0.6749, respectively, as shown in Fig. 3b and d. Among the models tested, the model utilizing a transfer learning approach within the ResNet-152 framework demonstrates superior overall performance compared with the other three models.

Model accuracy performance for subordinate classification tasks

To provide a more comprehensive assessment of model performance for complex tasks, we analyze the effectiveness of the four frameworks for subordinate classification tasks. This analysis is performed for both the validation and test sets, and the resulting performances are shown in Fig. 4a and b. Within the validation set, the ResNet-152, VGG-16, DenseNet-201 and GoogLeNet frameworks have top-1 Acc values of 0.4000, 0.3818, 0.3742 and 0.3792, respectively, as shown in Fig. 4a. In our evaluation of the ResNet-152 framework, we compute the microaveraged and macroaveraged performance metrics. The microaveraged recall is 0.7917, and the macroaveraged recall is 0.3850, as shown in Fig. 4b. In contrast, within the test set, we observe varying degrees of performance degradation for the ResNet-152 (0.3633), VGG-16 (0.3674), and DenseNet-201 (0.3616) frameworks, as shown in Fig. 4a. Notably, the GoogLeNet model increased the top-1 Acc by approximately 0.02 units. The best performing frameworks in terms of microaveraged recall and macroaveraged recall performance are the DenseNet-201 and GoogLeNet frameworks, respectively (see Fig. 4b). Figure 4c compares the top-1 Acc performances of four transfer learning models, each incorporating a different core framework, for superordinate and subordinate classification tasks. For the validation set, we calculated the top-1 Acc at 30 epochs after convergence for each model and performed a significance test to evaluate the performance differences among various groups. The p values of 2.59×10^{-9}, 7.30×10^{-9}, and 4.21×10^{-11} highlight substantial statistical disparities. The performance metrics are depicted through scatter plots and error bars, with the average of all values shown by the height of the bar chart (see Fig. 4c).

For the test set, we apply a similar approach where the performance of each model is measured at 30 epochs following convergence. The significance test revealed significant differences, with p values of 5.28×10^{-10}, 5.37×10^{-10}, and 5.05×10^{-10}. These results emphasize the distinct capabilities of the models when dealing with the complexities inherent in superordinate versus subordinate classification tasks. As with the validation set, we visualize the performance data via scatter plots and error bars, summarizing the average performance in bar charts (see Fig. 4c).

Variation in evaluation metrics for subordinate classification tasks

Given the high-dimensional nature of subordinate classification tasks involving up to 41 categories, relying solely on the top-1 Acc metric may not adequately represent the ability of the model to handle complex tasks with similar categories. Therefore, we also include the top-5 Acc as an evaluation metric for this task. This approach provides a more comprehensive assessment of model performance, aligning better with the high noise and mixed diagnosis realities encountered in clinical scenarios. We construct four transfer learning models by replacing the core framework and analyzed their top-1 Acc and top-5 Acc for subordinate classification tasks in both the validation and test sets, as shown in Fig. 4d. In the validation set, the ResNet-152, VGG-16, DenseNet-201 and GoogLeNet frameworks achieve top-5 Acc values of 0.6698, 0.6579, 0.6884 and 0.7039, respectively. These values significantly exceed the top-1 Acc by more than 0.1 units. Within the test set, the performances of the ResNet-152, VGG-16, DenseNet-201 and GoogLeNet frameworks yield top-5 Acc values of 0.6760, 0.6546, 0.6582 and 0.6640, respectively. Notably, these values are lower than the performance metrics of the models within the validation set.

For the validation set, our detailed comparative analysis reveals that the model employing transfer learning with the GoogLeNet framework outperformed the other models in terms of the top-5 Acc. To assess the performance differences among the groups, we perform a significance test on the top-5 Acc data from 30 epochs after convergence. This test reveals significant disparities in the overarching classification task, notably showing the superiority of GoogLeNet over models based on the ResNet-152, VGG-16, and DenseNet-201 frameworks. The evidence shows strikingly low p values of 8.17×10^{-12}, 2.79×10^{-11}, and 7.65×10^{-8}.

Similarly, the model built on the ResNet-152 framework achieves the highest Acc in the test set comparison. Further significance testing on postconvergence top-5 Acc data highlights marked differences in performance

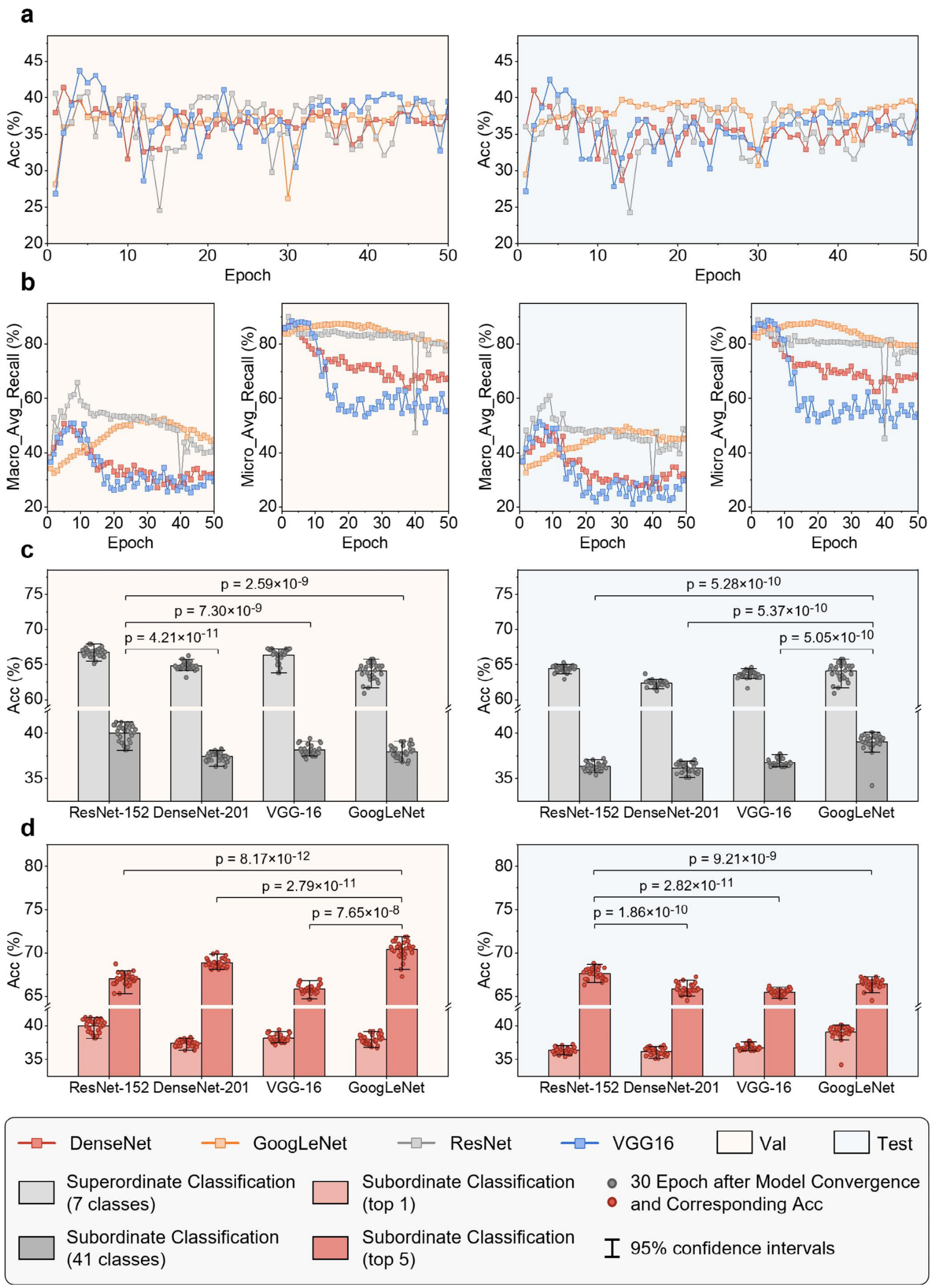

Fig. 4 Performance evaluation of the models trained on the respective core frameworks to verify the top-1 Acc (**a**), macroaveraged recall and microaveraged recall (**b**) for the subordinate classification tasks on the validation and test sets, comparative analysis and significance test to determine the performance differences between the top-1 Acc values for the superordinate classification tasks and the subordinate classification tasks on the validation and test sets (**c**), and comparison and significance test to determine the performance differences between the top-1 Acc and top-5 Acc for the subordinate classification tasks (**d**)

among groups for superordinate classification tasks. Notably, replacing ResNet-152 leads to significantly better performance than replacement with the GoogLeNet, DenseNet-201, and VGG-16 frameworks, as evidenced by the p values of 9.21×10^{-9}, 2.82×10^{-11}, and 1.86×10^{-10}, respectively.

Performance metric analysis is further carried out through scatter plots and error bars that effectively show result variability and reliability. These visual aids complement our findings by displaying average performances for models through bar chart heights in corresponding figures.

Data enrichment processing

To optimize the performance, we apply enhancement procedures such as flipping and rotating to the datasets. Transfer learning models are then constructed via four replacement core frameworks. The performance of these models is analyzed on both the superordinate and subordinate classification tasks within the validation and test sets following the data enhancement process. Figure 5 presents a performance analysis of the transfer learning models using a replacement ResNet-152 framework within the validation set and the test set before and after data enhancement (the performance comparison graphs for the remaining three models are included in Figures S2 to S4.). For the superordinate classification tasks, the ResNet-152 framework yields accuracy performances of 0.6599 and 0.6156 (see Fig. 5a) in the validation and test sets, respectively, after data enhancement. The validation sets for ResNet-152 yield microaveraged precision, microaveraged recall, macroaveraged precision, macroaveraged recall, microaveraged F1 and macroaveraged F1 of 0.4974, 0.8405, 0.6247, 0.4127, 0.6208, and 0.4811, respectively. The corresponding test sets have values of 0.4783, 0.8239, 0.6049, 0.3923, 0.6308 and 0.4720, as shown in Fig. 5b. The ResNet-152 framework yields top-1 Acc values of 0.3633 and 0.3950 (see Fig. 5c) and top-5 Acc values of 0.6309 and 0.6662 (see Fig. 5d) in the validation and test sets, respectively, for the subordinate classification tasks following data enhancement.

The use of data from enhanced processed datasets in both the superordinate and subordinate classification tasks does not significantly improve the overall performance. This suggests that conventional data enhancement strategies are of limited effectiveness for dental category medical X-ray data. Consequently, there is a need for targeted enhancement strategies to effectively improve the performance of models dealing with such specialized datasets.

Construction of an online web download platform

Our aim is to create a web-based dental research community equipped with libraries, datasets and communication facilities, as shown in Fig. 6a. Therefore, we create an online platform [27] to facilitate the downloading of our dataset content. Additionally, we include links to five other publicly available datasets and more than 20 links to relevant literature on the website. This aggregation helps to conveniently provide a comprehensive understanding of dental datasets. Figure 6b outlines the comprehensive development process of the website, which primarily involves front-end technologies such as hypertext markup language (HTML), cascading style sheets (CSS) and JavaScript. HTML is utilized to define the page structure and CSS styles of the presentation layer. JavaScript manages interactivity and dynamic behavior on the page. In terms of backend development, Spring Boot provides a productive and straightforward framework for building Java-based applications. My structured query language (MySQL) serves as the backend database to ensure reliable data storage and management. MyBatis-Plus extends MyBatis, increasing development efficiency and streamlining data manipulation procedures. This information can serve as a useful guide for developers. The design and functionality of our community interface are illustrated in Fig. 6c and e.

Discussion and conclusion

AI coupled with dental datasets has significantly advanced digitization in dentistry, with immense potential for clinical applications in automated diagnosis, disease prediction and taxonomic identification [34]. Dental datasets provide a valuable repository of prior information essential for implementing precision dentistry strategies. They encompass diverse research findings, such as clinical examination data, nested case designs, and cross-sectional designs of patient populations [15, 35]. However, importantly, the development and application of these dental datasets are still in their infancy, and the datasets currently available in oral health have significant limitations in terms of both scope and size. This situation hinders its potential utility as a reference point for comparing the performances of different models [19]. In addition, much of the relevant research is based on the application of a single model, which is focused primarily on diagnosing and assessing a single condition [12, 20, 21], limiting its widespread clinical use. Moreover, many datasets lack dedicated websites or community support [22–24], hindering their use and validation by other researchers. This limitation impedes progress toward relevant competitions and academic community development in this domain.

To address the shortcomings of existing databases, we created a new local oral dataset named LMCD-OR by compiling oral X-ray images from patients in local hospitals, offering a volume of data that surpasses many current collections. Most existing databases are limited in

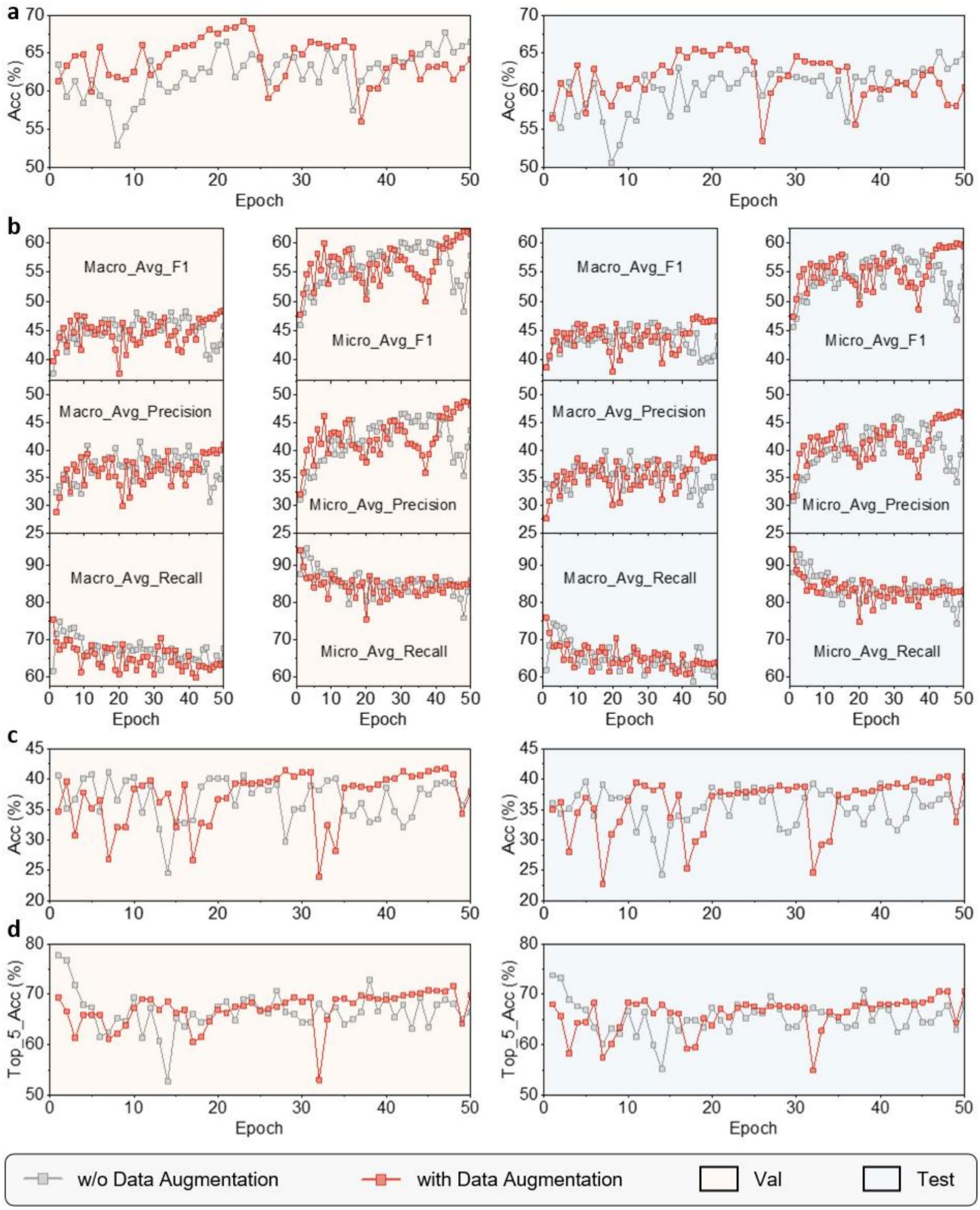

Fig. 5 Performance evaluation of the transfer learning model implemented via the ResNet-152 framework before and after data augmentation, including a comparison of the top-1 Acc values for the superordinate classification task on the validation and test sets (**a**), macroaveraged F1, microaveraged F1, macroaveraged precision and microaveraged precision, macroaveraged recall and microaveraged recall (**b**), comparison of top-1 Acc values for the subordinate classification task (**c**), and comparison of top-5 Acc values (**d**)

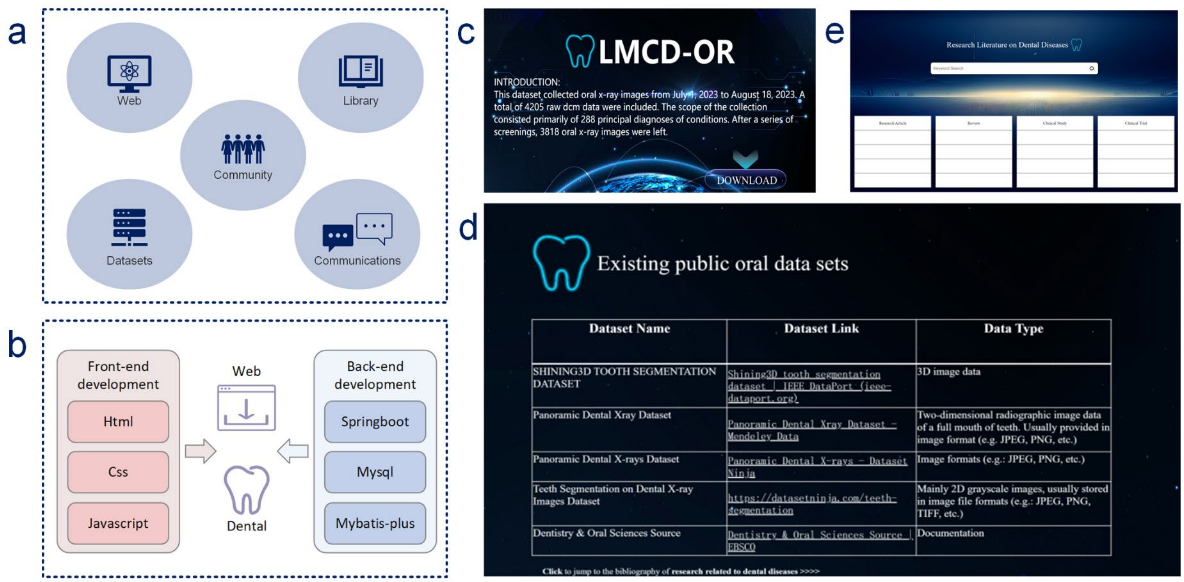

Fig. 6 Overview of the oral dataset resource platform, showing various components of the site: (**a**) the community structure, (**b**) the comprehensive development process of the website, (**c**) the download interface for accessing the dataset, (**d**) a webpage that provides links to established oral datasets, and (**e**) a literature search webpage designed specifically for the oral dataset

providing secondary diagnoses for specific conditions and lack information on broader classification levels [34–36]. In contrast, LMCD-OR boasts detailed hierarchical annotations and high labeling precision, encompassing both primary-level categories for oral outpatient visits and secondary-level diagnoses for specific diseases. This richness allows the use of a multiclass classification model that can learn various oral diseases simultaneously from a single X-ray image. Additionally, to ensure the ethical credibility of the research, all data included in LMCD-OR has been fully anonymized and de-identified. As detailed in Sect. 2.7, all patient identifiers were removed prior to data compilation, and strict protocols were followed to prevent any traceability back to individual patients. This guarantees that the dataset complies with ethical standards and legal regulations, ensuring full data anonymity and non-identifiability.

Moreover, LMCD-OR is a multilevel categorized diagnostic dataset for oral radiography, which addresses the limitations of traditional datasets that typically focus on single disease types. By providing both primary-level classifications for general dental conditions and secondary-level diagnoses for specific diseases, LMCD-OR supports multi-class disease classification. This multilevel annotation allows the simultaneous detection of multiple abnormalities and various disease types within a single scan, significantly enhancing its clinical relevance. With this structure, LMCD-OR facilitates the development of advanced classification models capable of diagnosing a broad range of oral diseases, setting a new benchmark for future research in oral disease diagnostics. The experimental outcomes underscore the advanced performance of LMCD-OR in oral image classification and disease diagnosis. Specifically, LMCD-OR can be leveraged in research projects to enhance AI-driven diagnostics, especially by supporting multimodal dataset integration studies. For instance, combining radiographic data with other clinical datasets, such as patient history or intraoral photography, can provide more comprehensive diagnostic insights. Such integration is anticipated to significantly improve the precision, scope, and reliability of clinical decision-making in diagnosing, treating, and forecasting oral health conditions.

Additionally, an independent web portal has been developed to provide seamless access to the LMCD-OR dataset for researchers. The platform offers key features designed to facilitate data access and enhance usability. These include a user-friendly interface for downloading the dataset, links to other established oral health datasets, and a literature search tool specifically tailored for oral radiography research. By offering these resources, the platform aims to foster collaboration and streamline the process for researchers to explore, download, and utilize LMCD-OR in their studies.

In this research, we employ a model skeleton replacement strategy aimed at enhancing the architecture and optimizing the application of four key visual neural network classification models. To comprehensively evaluate the performance of these models across a variety of classification tasks, we rely on two critical metrics: the top-1

Acc and the top-5 Acc. For a more comprehensive evaluation of the model, we also use microaveraged precision, microaveraged recall, microaveraged F1, macroaveraged precision, macroaveraged recall, and macroaveraged F1.

Focusing on the broader classification objectives, implementing the ResNet-152 framework delivers a top-1 Acc of approximately 0.7. This performance underscores the ability of the dataset to significantly enhance model efficiency. However, when more intricate classification tasks within the validation set are explored, the top-1 Acc for ResNet-152 decreases to only 0.4 in instances without data augmentation techniques. This discrepancy demonstrates how incorporating more sophisticated classification goals substantially increases computational demands on the model, indicating considerable room for further refinement.

Incorporating the top-5 Acc alleviates the challenges created by noise and the inherent similarities among various detailed classification targets found in real-world scenarios. In this respect, ResNet-152 achieves a top-5 Acc of 0.6698 with the validation settings. While this demonstrates commendable performance, it also identifies avenues for potential enhancements in optimizing the model.

A comparative analysis among various model architectures shows that employing a baseline model with ResNet-152 leads to a marginal performance improvement of approximately 0.02 over other architectures, such as DenseNet-201, VGG-16, and GoogLeNet, for most evaluated tasks. Although this improvement may appear slight and statistically nonsignificant, it demonstrates the adaptability and relative strengths of our dataset compared with those of multiple models in diverse contexts.

In our study, we utilize traditional visual data augmentation techniques such as image flipping and contrast adjustment. Despite these strategies not yielding substantial performance gains for ResNet-152, possibly because the intrinsic model design mitigates these visual modifications, they remain essential for validating the utility of our dataset. The results show that data enhancement does not significantly improve overall performance. This suggests that conventional data enhancement strategies for dental category medical X-ray data have limited effectiveness and cannot adequately address the complexity and diversity of the data.

To further improve model performance, future research should consider the use of broader and more targeted data enhancement strategies. In particular, strategies that capture and enhance features of classification labels prove to be more effective in improving the model's classification ability. For example, enhancement methods based on domain knowledge in dentistry, such as modeling common oral lesion features or enhancing lesion details, prove to be more effective than traditional methods are. In conclusion, optimizing and adapting data enhancement strategies is an important way to unlock the full potential of our dataset and improve the overall performance of the model.

While our study enhances the utility of oral datasets, certain limitations are acknowledged. LMCD-OR comprises only intraoral radiographs from a single healthcare provider, which may limit the generalizability of the findings to diverse populations. The single-source nature of the dataset means that it may not fully capture the variability in oral health conditions across different demographics or clinical settings. Future expansions of the dataset will need to address several challenges, including the incorporation of data from diverse populations and the standardization of imaging quality. Variations in equipment and operator techniques can impact the clarity and angulation of images, which may introduce biases. To mitigate these issues, collaborations with multiple healthcare providers and the implementation of standardized imaging protocols will be essential for ensuring consistency and improving the dataset's applicability to a broader population. Additionally, the integration of gene/protein signaling networks through ODE-based theoretical modeling has proven to be crucial in disease prediction, as demonstrated in recent studies [37, 38]. Future research can explore the application of these models in conjunction with LMCD-OR to enhance oral disease prediction capabilities and broaden the scope of diagnostics.

Despite these limitations, our research advances AI analysis of oral X-ray images and lays a foundation for developing AI-assisted oral X-ray image analysis systems. Our findings set the stage for deeper integration and application of AI alongside big data analytics within dentistry.

LMCD-OR is a comprehensive repository that covers a wide range of clinical data on prevalent oral diseases and includes samples of rare conditions such as acute pericoronitis [39]. Our forward-looking plans involve crafting a competitive model leveraging this dataset, which will incorporate small-sample learning techniques to handle these less common clinical cases. This ambitious model has the potential to significantly contribute to creating a thorough and precise framework for the differential diagnosis of oral diseases [40, 41].

To foster cooperative research and knowledge exchange, we have launched a dedicated website to establish it as a cornerstone for the global oral disease research community. This platform currently facilitates data access and downloads, and we have included detailed access and usage policies to ensure transparency and credibility. Potential users can apply for access by submitting a formal request through the website, and access is granted to academic and clinical researchers following a brief review process. Usage is restricted to non-commercial research purposes. Additionally, we are committed to maintaining and updating the dataset. New images and labels will be periodically added, and version maintenance will ensure the ongoing validity of the existing data. We anticipate that this platform will evolve into an active research ecosystem featuring forums and spaces for dialogue, allowing researchers, clinicians, and data scientists to share insights, discuss discoveries, and generate new lines of inquiry.

To enhance the community aspects of our website, we encourage the initiation of diverse academic endeavors and collaborative projects, all united by the overarching aim of driving innovation in dental disease diagnostics. Our vision is for this open-access platform to become an indispensable tool in advancing oral disease research worldwide. Additionally, we will implement a user feedback mechanism to continuously improve the usability of the LMCD-OR dataset. Researchers using the dataset will be able to provide feedback through the platform, which will be regularly reviewed and addressed by our team. This process ensures that user suggestions and issues are considered in future updates, further fostering community engagement and improving the dataset's functionality.

Abbreviations

AI	Artificial intelligence
DICOM/DCM	Digital Imaging and Communications in Medicine
TDD	Tufts Dental Database
ODSI-DB	Oral diseases and sciences image database
SKaPa	Swedish Quality Registry for Caries and Periodontal Diseases
NDPBRN	National Dental Practice-based Research Network
QCR	Queensland Cancer Registry
AUTH	Authority
HKU/HA HKW IRB	University of Hong Kong/Hospital Authority Hong Kong West Cluster
CNN	Convolutional neural network
R-CNN	Region-based convolutional neural network
CBCT	Cone beam computed tomography
SQUIRE	Standards for quality improvement reporting excellence
PNG	Portable network graphics
Top-1	Acc-Top-1 accuracy
Top-5	Acc-Top-5 accuracy
SGD	Stochastic gradient descent optimizer
HTML	Hypertext markup language
CSS	Cascading style sheet
MySQL	My structured query language
HIV	Human immunodeficiency virus

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12967-024-05741-3.

Supplementary Material 1

Acknowledgements

We extend our gratitude for the insightful discussions with Hao-Man Chen, Fan-Xuan Chen, Qing Wang, and Yu Xia. We would like to express our special appreciation to Professor Jian-Wei Shuai for his invaluable assistance in structuring the manuscript, conducting the literature review, and synthesizing the findings. Additionally, we acknowledge the meticulous guidance provided by Xue-Yin Bao regarding the illustrations created for this study.

Author contributions

J.Q.Z.: Data curation, Investigation, Writing - original draft. L.Z.: Data curation, Investigation, Writing - original draft. Z.F.M.: Data curation, Investigation, Writing - original draft. L.H.C.: Data curation, Investigation, Writing - original draft. Y.C.W.: Investigation, Methodology, Visualization. L.H: Conceptualization, Funding acquisition, Project administration, Supervision, Writing - review and editing. Q.Z.: Conceptualization, Funding acquisition, Project administration, Supervision, Writing - review & editing. F.F.S.: Conceptualization, Funding acquisition, Project administration, Supervision, Writing - review and editing.

Funding

This study is supported by Natural Science Foundation of Liaoning Province (Grant No. 2023-MS-288), Fundamental Research Funds for the Liaoning Universities (Grant No. LJ212410146026), the Ministry of Education Industry-Academia Talent Development Program (Grant No. 202101160011), and the Key Program of Translational Medicine Fund of Wenzhou Research Institute of Shanghai University (Grant No. SDTMF2023KP04).

Data availability

LMCD-OR is freely available at http://dentaldataset.zeroacademy.net/.

Declarations

Ethics approval and consent to participate

This study was approved by the Medical Ethics Review Committee of the Affiliated Stomatological Hospital of Wenzhou Medical University (Ethics No. WYKQ2023010). Due to the retrospective nature of the study and the use of anonymized, nonidentifiable radiographs, the requirement for informed consent was waived by the institutional review board.

Consent for publication

Not applicable.

Competing interests

The authors declare that the research was conducted without any commercial or financial relationships that could be construed as potential conflicts of interest.

Received: 23 August 2024 / Accepted: 3 October 2024
Published online: 14 October 2024

References

1. Spanemberg JC, Cardoso JA, Slob EMGB, et al. Quality of life related to oral health and its impact in adults. J Stomatology Oral Maxillofacial Surg. 2019;120(3):234–9.
2. Jain N, Dutt U, Radenkov I, et al. WHO's global oral health status report 2022: actions, discussion and implementation. Oral Dis. 2024;30(2):73–9.
3. Tu C, Wang G, Hu Z, et al. Burden of oral disorders, 1990–2019: estimates from the global burden of Disease Study 2019. Archives Med Sci. 2023;19(4):930–40.
4. de Abreu MHNG, Cruz AJS, Borges-Oliveira AC, et al. Perspectives on Social and Environmental Determinants of Oral Health. Int J Environ Res Public Health. 2021;18(24):13429.
5. Zhu F, Shuai Z, Lu Y, et al. oBABC: a one-dimensional binary artificial bee colony algorithm for binary optimization. Swarm Evol Comput. 2024;87:101567.
6. Liu L, Wei Y, Zhang Q, et al. SSCRB: Predicting circRNA-RBP interaction sites using a sequence and structural feature-based attention model. IEEE J Biomedical Health Inf. 2024;28(3):1762–72.
7. Yang X, Sun J, Jin B et al. Multi-task aquatic toxicity prediction model based on multi-level features fusion. J Adv Res, 2024; S2090-1232.
8. Zhu F, Niu Q, Li X et al. FM-FCN: A Neural Network with Filtering Modules for Accurate Vital Signs Extraction, Research, 2024; 7: 0361.
9. Jifeng W, Li Z, Jianqiang S, et al. Predicting drug-induced liver injury using graph attention mechanism and molecular fingerprints. Methods. 2024;221:18–26.

10. Zhu F, Ding J, Li X, et al. MEAs-Filter: a novel filter framework utilizing evolutionary algorithms for cardiovascular diseases diagnosis. Health Inform Sci Syst. 2024;12(1):8.
11. Gao H, Sun J, Wang Y, et al. Predicting metabolite-disease associations based on auto-encoder and non-negative matrix factorization. Brief Bioinform. 2023;24(5):bbad259.
12. Panetta K, Rajendran R, Ramesh A, et al. Tufts Dental Database: a Multimodal Panoramic X-Ray dataset for Benchmarking Diagnostic systems. IEEE J Biomedical Health Inf. 2022;26(4):1650–9.
13. Hwang JJ, Jung YH, Cho BH, et al. An overview of deep learning in the field of dentistry. Imaging Sci Dentistry. 2019;49(1):1–7.
14. Khan R, Akbar S, Khan A, et al. Dental image enhancement network for early diagnosis of oral dental disease. Sci Rep. 2023;13(1):5312.
15. Finkelstein J, Zhang F, Levitin SA, et al. Using big data to promote precision oral health in the context of a learning healthcare system. J Public Health Dent. 2020;80:S43–58.
16. Tan JY, Adeoye J, Thomson P, et al. Predicting overall survival using machine learning algorithms in oral cavity squamous cell carcinoma. Anticancer Res. 2022;42(12):5859–66.
17. Chau RCW, Li GH, Tew IM, et al. Accuracy of Artificial Intelligence-based photographic detection of Gingivitis. Int Dent J. 2023;73(5):724–30.
18. Kyventidis N, Angelopoulos C. Intraoral radiograph anatomical region classification using neural networks. Int J Comput Assist Radiol Surg. 2021;16(3):447–55.
19. Duncan WD, Thyvalikakath T, Haendel M, et al. Structuring, reuse and analysis of electronic dental data using the oral health and Disease Ontology. J Biomedical Semant. 2020;11(1):8.
20. Sivari E, Senirkentli GB, Bostanci E, et al. Deep learning in diagnosis of Dental anomalies and diseases: a systematic review. Diagnostics. 2023;13(15):2512.
21. Welch ML, McIntosh C, Traverso A, et al. External validation and transfer learning of convolutional neural networks for computed tomography dental artifact classification. Phys Med Biol. 2020;65(3):035017.
22. Rašić M, Tropčić M, Karlović P, et al. Detection and segmentation of Radiolucent Lesions in the Lower Jaw on panoramic radiographs using deep neural networks. Medicina. 2023;59(12):2138.
23. Vinayahalingam S, Berends B, Baan F, et al. Deep learning for automated segmentation of the temporomandibular joint. J Dent. 2023;132:104475.
24. Carvalho JS, Lotz M, Rubi L, et al. Preinterventional Third-Molar Assessment using Robust Machine Learning. J Dent Res. 2023;102(13):1452–9.
25. Park W, Schwendicke F, Krois J, et al. Identification of Dental Implant systems using a large-scale Multicenter Data Set. J Dent Res. 2023;102(7):727–33.
26. Alevizakos V, Bekes K, Steffen R, et al. Artificial intelligence system for training diagnosis and differentiation with molar incisor hypomineralization and similar pathologies. Clin Oral Invest. 2022;26(12):6917–23.
27. Dental Data Sets. http://dentaldataset.zeroacademy.net/ (2024), Accessed 3 Feb 2024.
28. Kaggle. https://www.kaggle.com/ (2010), Accessed 3 Feb 2024.
29. Al-Sarem M, Al-Asali M, Alqutaibi AY, et al. Enhanced tooth region detection using Pretrained Deep Learning models. Int J Environ Res Public Health. 2022;19(22):15414.
30. Sikha OK, Bharath B. VGG16-random fourier hybrid model for masked face recognition. Soft Comput. 2022;26(22):12795–810.
31. Lee JH, Kim DH, Jeong SN, et al. Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. J Dent. 2018;77:106–11.
32. Athisayamani S, Antonyswamy RS, Sarveshwaran V, et al. Feature extraction using a residual deep convolutional neural network (ResNet-152) and optimized feature dimension reduction for MRI brain tumor classification. Diagnostics. 2023;13(4):668.
33. Meng M, Zhang M, Shen D, et al. Differentiation of breast lesions on dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) using deep transfer learning based on DenseNet201. Medicine. 2022;101(45):e31214.
34. Bernauer SA, Zitzmann NU, Joda T. The Use and performance of Artificial Intelligence in Prosthodontics: a systematic review. Sensors. 2021;21(19):6628.
35. Gilbert GH, Fellows JL, Allareddy V, et al. Structure, function, and productivity from the National Dental Practice-Based Research Network. J Clin Translational Sci. 2022;6(1):e87.
36. Song IS, Shin HK, Kang JH, et al. Deep learning-based apical lesion segmentation from panoramic radiographs. Imaging Sci Dentistry. 2022;52(4):351–7.
37. Jun J, Fei X, Zhilong L, et al. Quantifying the underlying landscape, entropy production and biological path of the cell fate decision between apoptosis and pyroptosis. Chaos Solitons Fractals. 2024;178:114328.
38. Jin J, Xu F, Liu Z et al. Biphasic amplitude oscillator characterized by distinct dynamics of trough and crest. Phys Rev E 2023; 108(6 – 1):064412.
39. Hu H, Feng Z, Lin H, et al. Modeling and analyzing single-cell multimodal data with deep parametric inference. Brief Bioinform. 2023;24(1):bbad005.
40. Hu H, Feng Z, Lin H, et al. Gene function and cell surface protein association analysis based on single-cell multiomics data. Comput Biol Med. 2023;157:106733.
41. He Q, Zhong CQ, Li X, et al. Dear-DIA^{XMBD}: deep autoencoder enables deconvolution of Data-Independent Acquisition Proteomics. Research. 2023;6:0179.

Publisher's note