Computers in Biology and Medicine 192 (2025) 110279

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

NFR-EDL: Non-linear fuzzy rank-based ensemble deep learning for accurate diagnosis of oral and dental diseases using RGB color photography

Pouyan Razmjouei [a], Elaheh Moharamkhani [b], Seyed Sasan Aryanezhad [c], Mohammad Shokouhifar [d,**], Mehdi Hosseinzadeh [e,f,*], Behrouz Zadmehr [g]

[a] Department of Electrical & Computer Engineering, Valie-e-Asr University of Rafsanjan, Kerman, Iran
[b] IRO, Computer Science Department, University of Halabja, Halabja, Iraq
[c] Department of Oral & Maxillofacial Radiology, School of Dentistry, Islamic Azad University of Isfahan (Khorasgan), Isfahan, Iran
[d] DTU AI and Data Science Hub (DAIDASH), Duy Tan University, Da Nang, 550000, Viet Nam
[e] School of Computer Science, Duy Tan University, Da Nang, 550000, Viet Nam
[f] Jadara Research Center, Jadara University, Irbid 21110, Jordan
[g] Amirkabir University of Technology, Tehran, Iran

ARTICLE INFO

Keywords:
Oral health
Dental diseases
Convolutional neural networks
Deep learning
Ensemble learning
Fuzzy rank-based method

ABSTRACT

Background: Oral health plays a vital role in our daily lives, affecting essential activities like eating, speaking, and smiling. Poor oral health can lead to significant social, psychological, and physical consequences, which makes early and accurate diagnosis incredibly important. Recent advances in artificial intelligence (AI) are opening new doors in oral health care, offering faster, more accurate ways to identify dental issues and improve overall care.
Methods: This paper uses RGB color photography to introduce a non-linear Fuzzy Rank-based Ensemble Deep Learning model (NFR-EDL) for diagnosing oral and dental diseases. The model utilizes four deep Convolutional Neural Network (CNN) base models to analyze high-resolution color images of the oral cavity. The CNN base models are initially trained to generate confidence scores, which are subsequently mapped onto distinct functions with varying concavities, resulting in non-linear fuzzy ranks. Then, these ranks are combined into a final score to minimize the deviation from expected results. This method aims to provide accurate, reliable identification of oral and dental disease diagnosis by fusing many base models and considering uncertainty in decision-making while utilizing the rich visual information available in RGB images.
Results: The experimental results demonstrate that the proposed NFR-EDL model achieves accuracies of 97.08 %, 84.00 %, 89.86 %, and 94.66 % on the Kaggle, MOD, ODSI-DB, and OaDD datasets, respectively. These results demonstrate the model's exceptional accuracy and effectiveness in diagnosing oral and dental diseases, outperforming existing techniques and enhancing diagnostic reliability in clinical settings.
Conclusion: Deploying the NFR-EDL model in clinical settings offers a highly accurate and reliable tool for diagnosing oral and dental diseases, enhancing early detection, personalizing patient care, and reducing diagnostic errors to ultimately improve patient outcomes and the efficiency of dental care delivery. This approach reduces uncertainty in decision-making, ensuring that diagnoses are made with high confidence.

1. Introduction

Oral disease is a significant global public health concern, particularly concerning common dental conditions such as periodontitis and dental caries [1]. Periodontitis, a chronic inflammatory disease of the gums and teeth, is characterized by the breakdown of alveolar bone and surrounding tissues, particularly the periodontal ligament. Dental plaque is the primary cause of periodontitis, which initiates a cascade of inflammatory responses leading to the destruction of periodontal tissues. Research has indicated that periodontitis may be linked to other major systemic disorders, including an increased risk of cardiovascular disease [2]. Dental caries, caused primarily by acid erosion, can damage tooth structure, with intraoral bacteria producing the majority of the acid. Periodontitis and dental caries are prevalent oral disorders that

* Corresponding author.
** Corresponding author.
E-mail address: mehdihosseinzadeh@duytan.edu.vn (M. Hosseinzadeh).

https://doi.org/10.1016/j.compbiomed.2025.110279
Received 5 August 2024; Received in revised form 14 April 2025; Accepted 24 April 2025
Available online 3 May 2025

significantly impact people's quality of life [3]. These two conditions affect a substantial portion of the global population and can profoundly influence both general and oral health. Therefore, one of the most critical responsibilities of dentists is the prevention of dental cavities and periodontitis. Moreover, the early detection of these conditions has traditionally played a crucial role in their effective treatment [4].

Clinical and radiographic exams are the primary means of detecting dental caries and periodontitis. Traditional diagnostic methods such as clinical and radiographic examinations offer high accuracy but can be invasive and require frequent visits, which may cause patient discomfort. Recently, several researchers have proposed Artificial Intelligence (AI)-driven oral healthcare systems due to the potential benefits of AI [5–7]. The majority of these studies have used Dental Panoramic Radiography (DPR) to identify tooth types (tooth numbering), as well as to recognize ages and dental caries [8,9]. While radiography remains the gold standard for detecting dental hard tissue diseases, recent advancements in high-resolution digital photography and AI have opened up new possibilities for non-invasive, radiation-free screening methods. RGB color photography of the oral cavity offers several advantages, including the ability to capture surface-level changes, soft tissue abnormalities, and early-stage lesions that may not be visible on radiographs [10]. Moreover, color images provide rich visual information about tissue health, inflammation, and discoloration that can be crucial for early disease detection. Dentists usually give a complete diagnostic and treatment plan after evaluating different areas of a patient's oral health [11]. Generally, AI-assisted RGB color photography complements traditional diagnostic methods by detecting visible signs of periodontitis and dental cavities. This approach can save time, reduce the workload for dentists, and minimize patient discomfort caused by frequent examinations [12]. While dental periapical radiography remains essential for diagnosing certain dental hard tissue diseases, RGB-based methods offer a valuable additional screening tool that can be easily integrated into routine dental check-ups [13–15].

Deep learning, with the application of Convolutional Neural Networks (CNN), has cropped up as an immensely accurate means in dental radiography in terms of the identification and detection of various kinds of dental afflictions [16]. The use of such state-of-the-art technology has led to a significant augmentation of the area concerned with the application of deep learning in the field of dental disease detection. Despite the difficulties researchers face in this field, they are trying their best to overcome all such problems by devising new and unique methodologies. Consequently, deep learning has now become an extensively used tool in the field of medicine, where it is of great value in the pursuit of improved healthcare results [17].

While AI-based tools for dental health diagnostics have come a long way, there are still some significant gaps that need addressing. Many existing studies focus on using deep learning, particularly CNNs, with radiographic images like DPR to detect dental issues such as caries and periodontitis. However, this reliance on radiography brings certain drawbacks: it requires specialized equipment, exposes patients to radiation, and is not ideal for frequent screenings. Additionally, radiographs are great for viewing hard tissue but don't provide much information on soft tissues or early-stage issues that may be visible on the surface. Surprisingly, few studies have looked into using non-invasive RGB color photography, which captures surface details and soft tissue conditions that radiographs might miss. Moreover, most existing AI-driven models struggle with managing uncertainty in their predictions, which is crucial in clinical settings where even a small error can make a big difference. To help fill these gaps, this study introduces a hybrid fuzzy-ensemble deep learning model that utilizes RGB images, aiming to improve diagnostic accuracy while keeping uncertainty in check, ultimately making AI diagnostics more reliable for real-world dental care.

In this paper, we present a hybrid deep learning method based on fuzzy set theory and ensemble learning techniques. This method involves multiple pre-trained deep learning models as base classifiers integrated with custom deep-base learners. Our goal is to improve the accuracy and reliability of diagnosing oral and dental diseases while considering uncertainty in the final decision-making. Four base models are trained, and their confidence scores are obtained, which are then transformed using three non-linear functions to create fuzzy rankings. These confidence scores and non-linear fuzzy ranks are used to generate the final evaluation values. An overall score is computed to minimize deviances from predicted results. Notably, low deviances from the predicted results imply a higher level of confidence in identifying the disease. The proposed model is utilized to evaluate and diagnose oral and dental diseases using the Kaggle, MOD, ODSI-DB, and OaDD datasets.

Several innovations to improve the diagnosis of oral and dental diseases are presented in this paper, which we will describe below: First, our novel fuzzy rank-based ensemble method combined established networks with fuzzy logic, effectively addressing uncertainty in medical image classification and outperforming traditional ensembles. Second, an adaptive weighting mechanism dynamically adjusted the contributions of each base model based on their performance across various disease categories, resulting in a more robust classification system. Third, we introduced the OaDD dataset, which integrated multiple open-source datasets to provide a diverse benchmark for classification tasks. Finally, our extensive comparative analysis showed significant improvements in classification accuracy and generalization. While real-world clinical trials were still pending, our method's performance across various datasets suggested strong potential for clinical applications and enhanced interpretability compared to traditional black-box models, fostering trust in AI-assisted diagnoses. Together, these innovations established a more effective and reliable framework for diagnosing oral and dental diseases. Specifically, the main contributions of this paper are as follows.

- Introducing a Non-linear Fuzzy Rank-based Ensemble Deep Learning (NFR-EDL) model for the diagnosis of oral and dental diseases.
- Utilizing four deep feature extractors, including VGG16, DenseNet169, ResNet50, and SqueezeNet, for feature extraction and generation of initial classification scores.
- Implementing a fuzzy rank-based method to combine individual model outputs, enhancing the final classification accuracy and improving reliability.
- Extensive experimental results demonstrate that NFR-EDL can improve the classification of oral and dental images across multiple metrics.

The paper progresses as follows: Section 2 examines related research. Section 3 describes the proposed method, followed by the research findings in Section 4 and the conclusions in Section 5.

2. Related research

In this section, we explore studies focused on various methods for addressing dental and oral health issues, such as gingivitis, oral and mouth cancer, dental caries, dental fluorosis, periodontitis, tooth damage, dental calculus, plaque, and tooth loss. These techniques are categorized into radiographic imaging and RGB color photography.

2.1. Techniques based on radiographic imaging

Prajapati et al. [18] suggested enhancing the accuracy of automated medical assistance systems using a labeled dataset that includes 251 radiography X-ray images, divided into three classes, with the help of CNN for highlighting issues in medical imaging recognition, segmentation, and classification with high precision. They explored, in their study, the performance of CNN in diagnosing small labeled dental datasets and showed how transfer learning effectively enhanced accuracy. The results from experiments with three different CNN architectures were indeed very encouraging, showing that large sets are crucial in tackling natural image classification problems.

Zhang et al. [19] introduced deep learning techniques adopted in the medical field to detect and classify teeth from dental periapical radiographs. The process of recognizing 32 teeth positions in an X-ray image due to the large number of teeth in adults is complicated. Therefore, they introduced a method of labeling trees to assign multiple labels to each tooth and introduced a cascade network structure for automatic identification. Its approach targeted key strategies to improve the performance of both detection and classification against outstanding performance in managing such complicated cases as missing, decaying, and filled teeth. Experimental results showed that precision and recall reached 95.8 % and 96.1 %, respectively, proving a huge increase in the performance of this challenging task.

Li et al. [20] introduced the prevalence of apical lesions as a common dental disease. The challenges of current endodontic treatments rely on X-ray radiography, which can be extremely time-consuming and complicated, with different shooting angles and different doses of radiation. They developed a lesion area analysis model based on CNN to simplify diagnosis. The preprocessing of image data was performed using a Gaussian high-pass filter, followed by iterative thresholding of X-ray images for segmenting them into individual samples of teeth. They used 70 % of the image database for training and validating the model, reserving the remaining 30 % for testing and estimating the accuracy. Finally, the proposed CNN model succeeded in enabling the automated diagnosis of apical lesions.

Endres et al. [21] compared the diagnostic accuracy of 24 Oral and Maxillofacial (OMF) surgeons to that of a predictive deep learning algorithm, revealing that dentists frequently misdiagnosed periapical radiolucencies on panoramic radiographs, confusing them with infections, granulomas, cysts, or tumors. The algorithm surpassed 14 of the 24 OMF doctors, attaining an average precision of 0.60 and an F1 Score of 0.58. Although trained on a restricted dataset and clinically tested against recognized standards, the algorithm exhibited considerable potential to aid OMF surgeons in precisely identifying these disorders.

Mao et al. [22] have proposed one deep learning-based method using CNN for furcation defects detection in periapical radiographs based on a dataset of 300 images of both with and without Furcation Involvement (FI) periapical radiographs, which produced an accuracy rate as high as 95 %. The effective image-masking technique used enhanced the contrast between the FI symptoms and the surrounding areas, thereby allowing for effective highlighting of the regions of interest in CNN model training. On the other hand, up to 94.97 % overall accuracy was attained from the proposed segmentation algorithm, outperforming conventional methods significantly. This study will provide feasibility and effectiveness related to furcation defect detection using deep learning algorithms and also bring into the limelight the potential of AI-assisted dental diagnosis in clinical practice.

Shafiq et al. [23] highlighted that the prevalence of periodontal diseases accounts for 5–10 % of health costs in developing economies, according to the World Health Organization (WHO). These diseases are terrible due to bacterial infection that gives rise to inflammation. They tried to classify three varieties of dental diseases using the CNN model. They conducted the experiments by using different pre-trained CNN models in association with Support Vector Machine techniques on the locally developed dataset provided by the Department of Radiology, Bahawal Victoria Hospital, Bahawalpur, Pakistan. The dataset contained 1067 dental images representing oral cancer, broken teeth, and dental caries, which were carefully labeled with the help of an expert dental surgeon. This proposed 24-layer CNN model has achieved a classification accuracy of 95.34 %, reflecting the enhanced performance of the developed model as a diagnostic tool in dentistry for the above-mentioned dental conditions.

Liu et al. [24] proposed a method of detection that is efficient for common lesions in digital dental X-ray images using a deep convolutional neural network. The dataset used within their work consisted of 188 images previously diagnosed by dentists from Qilu Hospital of Shandong University, including conditions such as periapical periodontitis, dental caries, and periapical cysts, among other common dental diseases. The images and labels were fed into four CNNs-VGG-16, InceptionV3, ResNet-50, and DenseNet-121 for training. The mean accuracy of the classification by the trained models turned out to be 95.9 %. Among them, the accuracy from the model DenseNet-121 reached 99.5 %. This research points out that the strategy of CNNs for interpreting digital dental X-ray images is effective and precise, hence proving they will be useful tools in auxiliary diagnosis.

Deepak and Krishna Bhat [25] presented an optimized deep learning-based CNN designed for the multi-classification of ocular diseases, including cataracts and glaucoma. In their study, they optimized three pre-trained CNN models, SqueezeNet, Darknet-53, and EfficientNet-b0, by varying the batch sizes (6, 8, and 10) and different optimizer types (SGDM, RMSProp, Adam) to achieve maximum accuracy in detecting various ocular conditions. Among these models, Darknet-53 stood out with the highest accuracy of 99.4 % on a test sample of 1000 images. The comparative performance analysis revealed that SqueezeNet, Darknet-53, and EfficientNet-b0 achieved accuracy rates of 95 %, 99.4 %, and 90 %, respectively. These findings underscore the significant impact of batch size and optimizer type on the performance of CNN models in diagnosing ocular diseases.

Kokomoto et al. [26] presented a fully automated method for calculating dental age using panoramic radiographs combined with deep learning techniques. Their approach utilized a substantial dataset of 8023 panoramic radiographs for training, along with 18,485 single-root and 16,313 multi-root dental germ images for the EfficientNetV2 model. The results indicated high performance in dental germ detection, where the average precision attained is 98.26 %. Further, the classification in single and multi-root dental germs arrived at an accuracy rate of 98.46 % and 98.36 %, respectively, for their Top-3 accuracies. The weighted average method performed better than the other methods, with less than one developmental stage error in the dental age calculation.

2.2. Techniques based on RGB color photography

The literature on dental and oral disease diagnosis has largely focused on radiographic imaging, such as periapical radiographs and digital X-rays, due to their effectiveness in detecting hard tissue issues. However, with advancements in artificial intelligence and high-resolution imaging, alternative methods like RGB color photography are gaining traction. Unlike radiography, RGB imaging captures surface-level changes and soft tissue abnormalities, providing a more comprehensive view of oral health, including inflammation, early lesions, and color changes. This method offers a non-invasive, radiation-free option that is especially valuable for routine dental check-ups and early detection, making it highly accessible for telemedicine and remote monitoring. While radiography is still critical for detailed hard-tissue assessments, RGB-based techniques show significant promise as complementary tools, supporting timely, efficient, and holistic approaches to oral health care through AI-powered models. In the following, we review the existing RGB-based techniques.

You et al. [27] developed a deep learning-based AI model designed to detect dental plaque on primary teeth in children, utilizing RGB intraoral photos. The model was trained on a dataset of 886 high-resolution RGB images of primary teeth, which allowed it to learn to identify plaque effectively. To assess its clinical feasibility, the model's performance was compared with that of an experienced pediatric dentist, who manually examined the same images. Plaque detection accuracy was evaluated using the Mean Intersection-over-Union (MIoU) metric, and the results showed that the AI model outperformed the dentist's accuracy in both initial assessments and follow-up evaluations after one week. Importantly, the model demonstrated clinically acceptable performance, with no significant difference when compared to the dentist's manual diagnosis. This study highlights the potential of AI, particularly through the use of RGB images, in enhancing pediatric

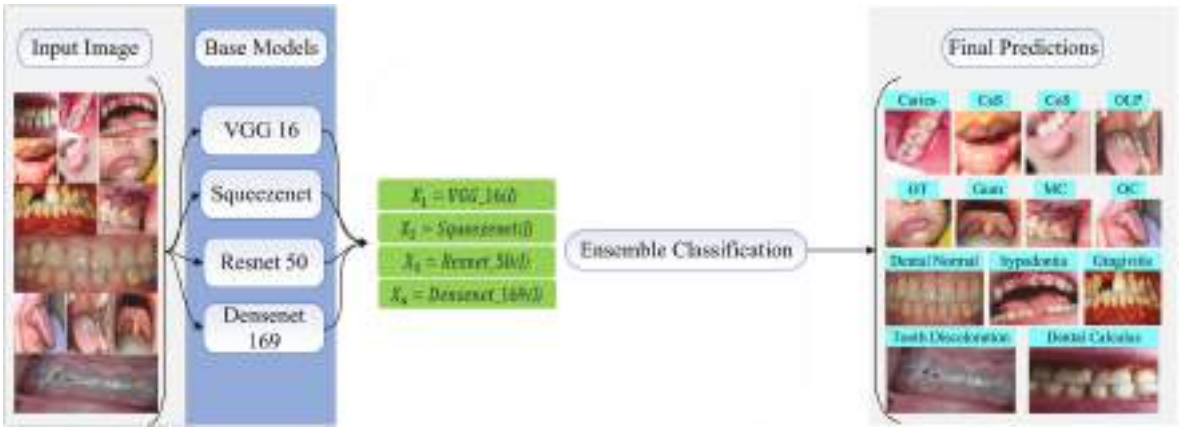


Fig. 1. Schematic diagram of the proposed NFR-EDL model.

oral health by offering a reliable, consistent, and non-invasive tool for plaque detection that could complement traditional diagnostic methods.

Alalharith et al. [28] developed a deep learning-based approach to help detect early signs of gingivitis in orthodontic patients by using Faster Region-Based CNN (Faster R-CNN) on intraoral RGB images. The study included a dataset of 134 RGB images, with 80 % used for training and 20 % for testing. The team built two models: one to detect teeth as a region of interest and the other to detect signs of gingival inflammation within that region. The teeth detection model performed with high accuracy, achieving perfect scores in accuracy, precision, and mean average precision (mAP). The inflammation detection model, while slightly lower, still achieved good results, with an accuracy of 77.12 %, a precision of 88.02 %, and an mAP of 68.19 %. This research demonstrates that deep learning on RGB images can effectively support non-invasive, early gingivitis screening, potentially aiding in early dental intervention and reducing the global impact of periodontal disease.

Li et al. [29] proposed a deep learning-based method for screening gingivitis, dental calculus, and soft deposits from RGB oral photos, aiming to provide a cost-effective solution for areas with limited medical resources. They developed a novel Multi-Task Learning CNN model that not only classifies the presence of these dental issues but also localizes their positions on the images. With a dataset of 625 patients, the model achieved strong classification performance, with Area Under the Curve (AUC) scores of 87.11 % for gingivitis, 80.11 % for dental calculus, and 78.57 % for soft deposits. The model also demonstrated moderate accuracy in localizing these issues, which adds an interpretative layer to the diagnosis. By outperforming general-purpose CNNs, this study highlights the potential of deep learning for large-scale, non-invasive dental screenings, making it a promising tool for improving public dental health.

Garg et al. [30] showed that oral diseases, such as periodontal issues and dental caries, affect billions of people worldwide. They presented a lightweight machine-learning model for calculus detection in RGB images, which can work efficiently on low-end devices. An accuracy of 72.73 %, comparable with state-of-the-art solutions, was achieved by the model implemented with a modified MobileNetV3-Small neural network. On the other hand, the ResNet34-based model performed with even better accuracy, rising as high as 81.82 %. Both models have been tested within a mobile application to give patients a chance to schedule an appointment much earlier and thus limit the number of serious cases of oral diseases.

Park et al. [31] discussed the quality of life related to oral health and also discussed that despite the immense use of artificial intelligence in this domain, studies using RGB images are scant. In their paper, they proposed a deep CNN model designed to classify teeth with periodontal diseases based on optical color images taken from the front of the mouth. This innovative model, which employs one-dimensional convolutions in parallel, outperformed conventional models such as ResNet152 by an impressive 11.45 %. Moreover, the proposed architecture significantly enhanced training performance, particularly in scenarios with limited training data. This study highlights the potential of optical color images for the detection of periodontal disease, paving the way for the development of a mobile oral healthcare system in the future.

Jiang et al. [32] highlighted the fact that dental caries is one of the most common oral diseases. They mentioned that despite that the diagnosis using deep learning methods has gained much attention, most of the earlier approaches suffered from the following problems: underutilization of the important features caused by background noise or interference of irrelevant information. They addressed these challenges by proposing a fine-grained RGB image classification framework, CariesFG, which is made of four major components: the Multi-Spectral Channel Attention Module, the Position Attention Module, the Discriminative Point Selection strategy, and the Graph Convolution and Aggregation module. This framework outperformed state-of-the-art methods in terms of accuracy, F1 Score, and specificity, showing the efficacy of the network in learning discriminative features and relevant parts, which is helpful for improved diagnosis of dental caries.

Sulochana and Sumathi [33] introduced the global health challenge of oral cancer, which faces difficulties in early detection due to many challenges related to the presence of various artifacts, noise, and poor visibility of lesions in conventional imaging techniques that include intraoral RGB photography. They have proposed an automated, adaptive, and multi-modal pre-processing pipeline to assist in improving the early detection of oral cancer using these imaging techniques. This is a high-performance pipeline that, with region-of-interest cropping, special noise filtering, and adaptive histogram equalization, significantly enhances image quality, reduces noise, improves the visibility of lesions, and increases the visibility of features. Among these various techniques, the Iterative Adaptive Weighted Median Filter (IAWMF) could provide maximum fidelity in denoising, while median filtering turned out to be useful in noise reduction, thus contributing to more accurate diagnostic performance.

Sampath et al. [34] developed a deep learning model called OralNet to help with non-invasive oral cancer detection using RGB images of the lips and tongue. The model uses a deep CNN based on ReNet-50 to extract important features from the images and then applies logistic regression to make the final diagnosis. After training the model on a benchmark RGB image dataset, it achieved an impressive accuracy of 97.8 % in identifying oral cancer, which is much higher than many other methods. This approach offers a promising, non-invasive alternative to traditional biopsy techniques, making it easier for doctors to detect oral cancer early and improve patient outcomes.

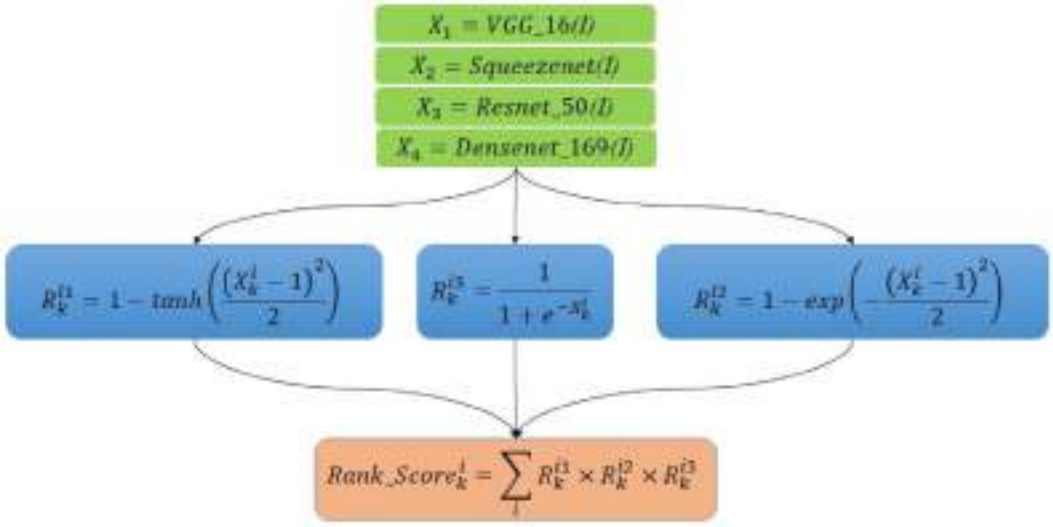


Fig. 2. Ensemble method using non-linear functions in the proposed NFR-EDL model.

3. Proposed NFR-EDL methodology

The paper introduces a novel NFR-EDL model for developing predictive models by using deep neural networks, ensemble learning, and a non-linear fuzzy rank-based method. This methodology entails using four basic learners, namely ResNet50, DenseNet169, SqueezeNet, and VGG16. The learners enhance the performance of the model via collaborative and coordinated efforts. The weights of the first layers of the base learners are pre-trained using the ImageNet dataset. This fine-tuning strategy balances transfer learning benefits with task-specific optimization [18,25]. Freezing all but the final five layers preserves generic low-level features while enabling adaptation to our datasets. We initialized base models (VGG16, DenseNet169, ResNet50, SqueezeNet) with pre-trained ImageNet weights, applied a low learning rate (1e-4) to prevent catastrophic forgetting, and kept batch normalization layers trainable for domain adaptation.

The diagram of the suggested NFR-EDL model is shown in Fig. 1. A technique is outlined for determining the fuzzy ranks of classes using non-linear functions. In this technique, the products of the ranks from the four base learners are computed, with the lower rank being selected

Table 1
Details of the utilized datasets.

Dataset	Reference	Class	Original Images	Train		Test (20 %)
				Original (80 %)	Augmented	
Kaggle	[35]	Calculus	1296	1037	2076	259
		Caries	2382	1906	3808	476
		Gingivitis	2349	1880	3760	469
		Hypodontia	1251	1001	2000	250
		Mouth Ulcer	2541	2033	4064	508
		Tooth discoloration	1834	1468	2934	366
MOD	[36]	CaS	79	64	180	15
		CoS	75	60	180	15
		Gum	61	49	144	12
		MC	90	72	216	18
		OC	54	44	120	10
		OLP	94	76	216	18
		OT	63	51	144	12
ODSI-DB	[37]	Calculus	146	116	360	30
		Normal	153	124	348	29
OaDD	[35],	Calculus	1442	1153	2436	289
		Caries	2382	1906	3808	476
		Gingivitis	2349	1880	3760	469
			1251	1001	2000	250
	[36],	Tooth discoloration	1834	1468	2934	366
		CaS	2504	2004	4004	500
		CoS	191	153	420	38
		Gum	61	49	144	12
	[37]	MC	90	72	216	18
		OC	54	44	120	10
		OLP	94	76	216	18
		OT	63	51	144	12
		Normal	153	124	348	29

Abbreviations.
CaS: chancre sore CoS: cold sore Gum: gingivostomatitis MC: mouth cancer.
OC: oral cancer OLP: oral lichen planus OT: oral thrush.

Fig. 3. Examples of images from the Kaggle dataset: A) Dental Tartar, B) Decay, C) Gingivitis, D) Hypodontia, E) Mouth Ulcers, and F) Tooth Discoloration [35].

as the predicted class. The application of non-linear functions contributes to the accuracy of the predictions and enhances the model's performance. Fig. 2 illustrates the Ensemble method, showcasing the output of the CNN model that is input to the non-linear functions. The proposed fuzzy rank-based group network, with four base learners, undergoes comparison using a variety of classification metrics. These metrics include Accuracy, Precision, Recall, and F1 Score.

The motivation behind converting classification probabilities into fuzzy numbers in our proposed NFR-EDL model stems from the need for a more nuanced, mathematically robust, and clinically relevant confidence measure. Traditional probability-based classification assumes a rigid certainty in outcomes, which is rarely the case in real-world medical diagnosis, wherein diseases often present overlapping symptoms. Fuzzy numbers offer a more flexible representation of this inherent uncertainty, closely mirroring the way clinicians assess diagnoses—considering degrees of membership rather than strict binary classifications. From a mathematical standpoint, fuzzy numbers provide a richer uncertainty model, allowing the incorporation of expert domain knowledge through well-defined membership functions. This enhances the interpretability of results while improving stability by reducing sensitivity to outlier classifiers. Clinically, this approach aligns with how medical decisions are made—by evaluating multiple interacting factors rather than relying on rigid categorical labels. Theoretically, our foundation in Zadeh's extension principle and possibility theory ensures a sound framework for integrating conflicting evidence from multiple classifiers, ultimately leading to a more reliable and insightful decision-making process in complex medical scenarios.

In the following, we detail the datasets used in this study in Subsection 3.1. Subsection 3.2 describes the architecture of the baseline CNN models. Subsection 3.3 introduces the proposed NFR-EDL model, and finally, Subsection 3.4 outlines the performance metrics.

3.1. Datasets

In this study, the fuzzy rank-based model is trained and validated using four datasets. A summary of data sources of oral and dental

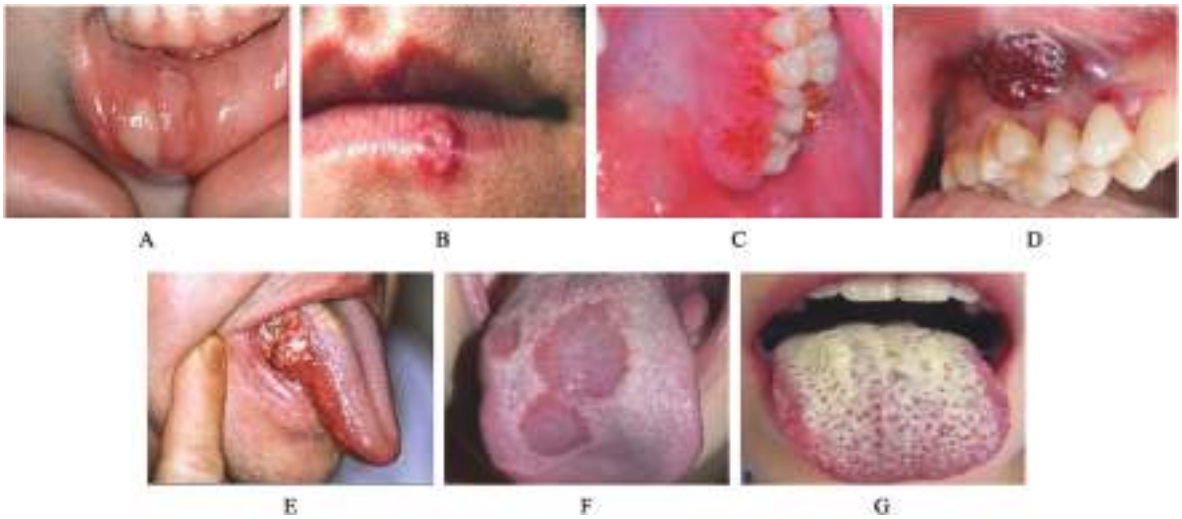


Fig. 4. Examples of images from the MOD dataset: A) Chancre Sore (CaS), B) Cold Sore (CoS), C) Gingivostomatitis (Gum), D) Mouth Cancer (MC), E) Oral Cancer (OC), F) Oral Lichen Planus (OLP), and G) Oral Thrush (OT) [36].

Fig. 5. Examples of images from the ODSI-DB dataset: A) Healthy (Normal), and B) Tartar or Inflammation (Calculus) [37].

diseases is given in Table 1. Each dataset is split into 80 % for training and 20 % for testing, ensuring that the distribution of classes is consistent across both sets. To further improve the model's performance and reliability, we use a 5-fold cross-validation approach, where the training set is divided into five separate folds. To further enhance the model's generalization, we apply data augmentation techniques such as rotation, scaling, zooming, and cropping, independently to each fold within the training set. Importantly, we ensure that each fold contains an even representation of all classes, so that the model can learn from a balanced dataset. Additionally, we ensure that during both the data splitting and augmentation process, no patient data is shared across folds, effectively preventing any risk of data leakage and maintaining the integrity of the evaluation process. This strategy helps the model to better handle real-world variability and improves its robustness in unseen data. The description of the four datasets can be summarized as follows:

Dental Status Dataset (Kaggle): The Dental Status dataset, available on Kaggle [35], is an open-source resource designed for dental research and analysis. This dataset is a comprehensive collection of images encompassing a wide range of dental conditions, such as decay, tartar, gingivitis, tooth discoloration, ulcers, and Hypodontia. The dataset comprises 11653 images, divided into six categories including 2382 images of decay, 1296 images of dental tartar, 2349 images of Gingivitis, 1834 images of tooth discoloration, 2541 images of ulcers, and 1251 images of Hypodontia. The images are in various formats, and all are converted to JPG format with a resolution of 600 × 600. Fig. 3 provides examples from each of the six categories.

Mouth and Oral Diseases (MOD) Dataset: The MOD dataset [36], includes images sourced from dental clinics in Okara, Punjab, Pakistan, as well as various dental websites. This dataset provides labeled images for seven types of oral and mouth diseases: Gingivostomatitis (Gum), Chancre Sore (CaS), Cold Sore (CoS), Oral Lichen Planus (OLP), Oral Thrush (OT), Mouth Cancer (MC), and Oral Cancer (OC). Importantly, no personal details like age, gender, or height were collected during the photography process. Each disease category in the dataset was carefully labeled by specialist dentists, resulting in a collection of 516 images across seven categories: 61 images of Gingivostomatitis, 79 of Chancre Sore, 75 of Cold Sore, 94 of Oral Lichen Planus, 63 of Oral Thrush, 90 of Mouth Cancer, and 54 of Oral Cancer. The images are in JPG format with a resolution of 224 × 224, and Fig. 4 presents some examples from each disease category.

Oral and Dental Spectral Image Database (ODSI-DB) Dataset: The third dataset utilized is ODSI-DB [37]. This dataset comprises 316 spectral images of the mouth and teeth, with 215 cases manually labeled using binary masks. The images encompass 35 different classes and are captured with two spectral cameras possessing distinct features. For classification purposes, this dataset is divided into two categories: images of healthy teeth (Normal) and images of teeth with tartar or inflammation (Calculus). The dataset includes 299 images, divided into two categories containing 146 images of tartar or inflammation (Calculus) and 153 images of healthy teeth (Normal). The images are in JPG format with a resolution of 640 × 640. Fig. 5 provides examples from each of the two categories.

Oral and Dental Diseases (OaDD) Dataset: The fourth dataset, referred to as OaDD, is a compilation of the previous three datasets and is categorized into 13 classes for classification purposes. The ulcer data category, sourced from the Dental Status dataset available on the Kaggle website, comprises 2541 images and is bifurcated into two categories: CaS and CoS. These categories are subsequently incorporated into the MOD dataset. Conversely, the tartar or inflammation (Calculus) data category from the ODSI-DB dataset, which includes 146 images, is amalgamated with the dental tartar data category in the Dental Status dataset from the Kaggle website. The remaining data categories are added to this dataset as new categories from each dataset. This not only enhances classification but also results in a comprehensive dataset for oral and dental diseases.

3.2. Deep CNN base learners

In this section, a method for identifying oral and dental diseases based on the fuzzy method is summarized. The core concept of this method is the utilization of several custom base learners, each customized with pre-trained models in ImageNet [38], to generate a confidence score for the presence of oral and dental diseases. The use of multiple base learners enhances the accuracy and reliability of disease diagnosis. Subsequently, these confidence scores are fused using the fuzzy method, taking into account uncertainty and diversity. The final diagnosis of diseases is based on the fuzzy set score. The objective of this method is to enhance the accuracy and reliability in diagnosing oral and dental diseases by leveraging the power of multiple base learners and considering uncertainty in the final decision. To implement this method, four base learners are utilized, and the method is evaluated on datasets related to oral and dental diseases. Initially, the base learners are trained, and confidence scores are obtained. These scores are then plotted on three different functions with varying concavities to create non-linear fuzzy ranks. By combining these three ranks, a composite score or final score is derived that aids in reducing the overall deviation from the expected case. A lower deviation indicates higher confidence in a specific disease. The disease exhibiting the smallest deviation is regarded as the winner and is assigned as the final value for the condition.

The following are brief details about the used base learners within the NFR-EDL model:

VGG16 (Visual Geometry Group 16): This model is a notable member of the CNN family. The Visual Geometry Group designs it from the University of London [39]. VGG16 utilizes 16 deep layers, enabling it to effectively extract features from images. The employment of multiple fully connected channels allows for the operation within a smaller and more optimal weight space for each filter. By leveraging high-depth and small filters, it can extract detailed information from images. The high efficiency of this model in pattern recognition has led to its widespread use in CNN architectures. In this model, a custom Fully

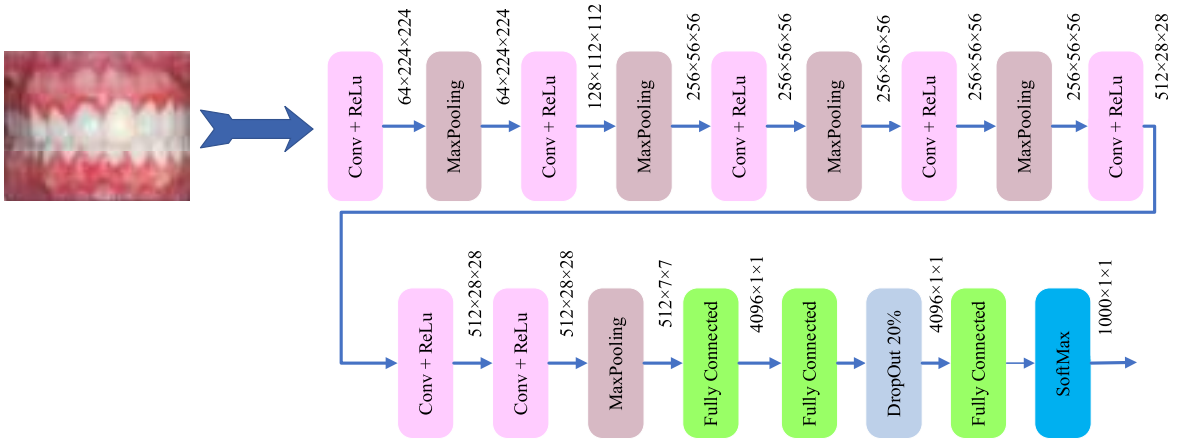


Fig. 6. Architecture of the VGG16 model.

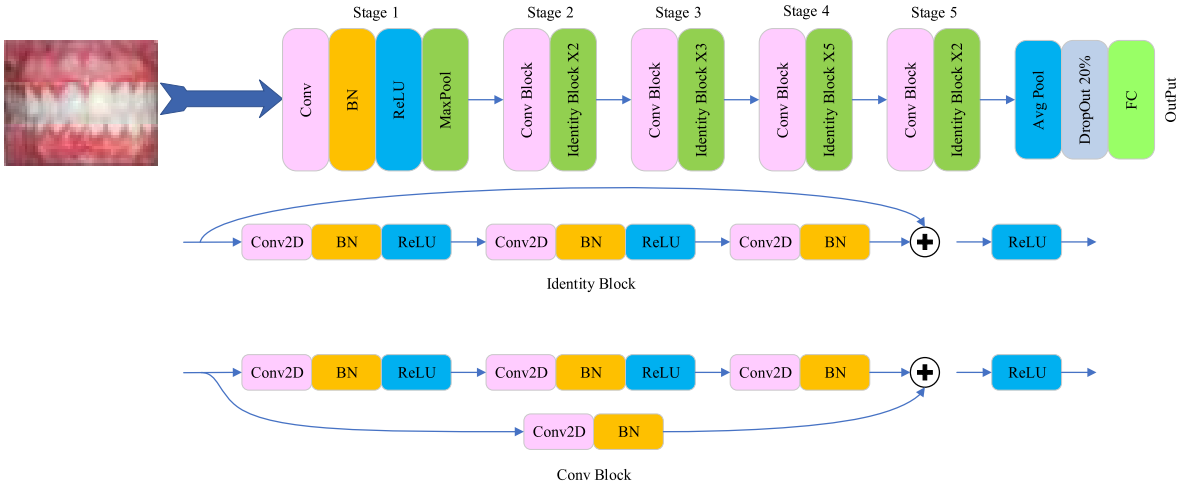


Fig. 7. Architecture of the ResNet50 model.

Connected layer with a Rectified Linear Unit (ReLU) activation function is added for accelerated learning, along with a DropOut layer of 20 % to prevent overfitting, based on the model structure. The architecture diagram of the VGG16 model is depicted in Fig. 6.

ResNet50 (Residual Network 50): This model was introduced in 2015 by researchers from Microsoft's research team [40]. The design of this model incorporates additional blocks known as Residual Blocks. These blocks enable the network to utilize residual information, thereby optimizing the performance for feature identification. ResNet50 is recognized as one of the most effective and reliable architectures in the field of image recognition. In this model, a custom DropOut layer of 20 % is incorporated to prevent overfitting. The architecture diagram of the ResNet50 model is depicted in Fig. 7.

DenseNet169: The DenseNet architecture was presented in 2017 by Gao Huang and his team at the Catholic University of Leuven [41]. This architecture is predicated on the direct connection of sequential layers to each other. In DenseNet, each layer is linked to all preceding layers and conveys information to subsequent layers without any intermediaries. This direct connection facilitates the efficient dissemination of information within the network. Moreover, the number of

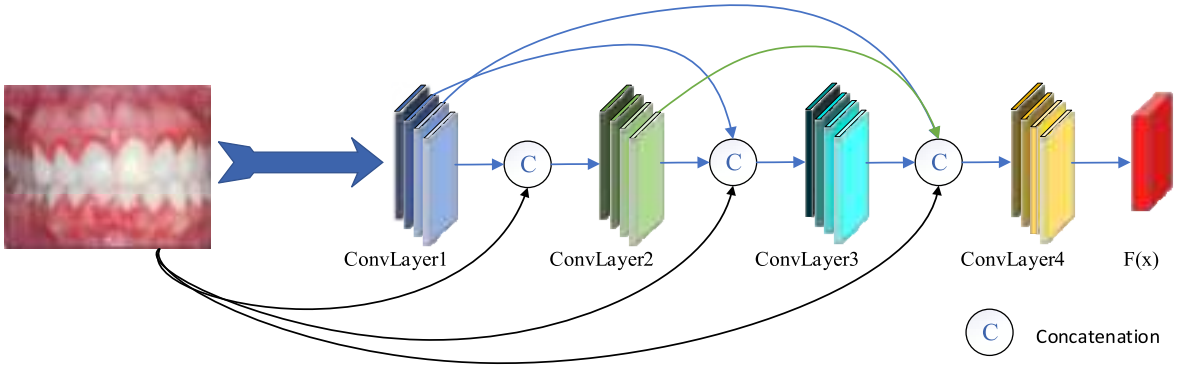


Fig. 8. Architecture of the DenseNet169 model.

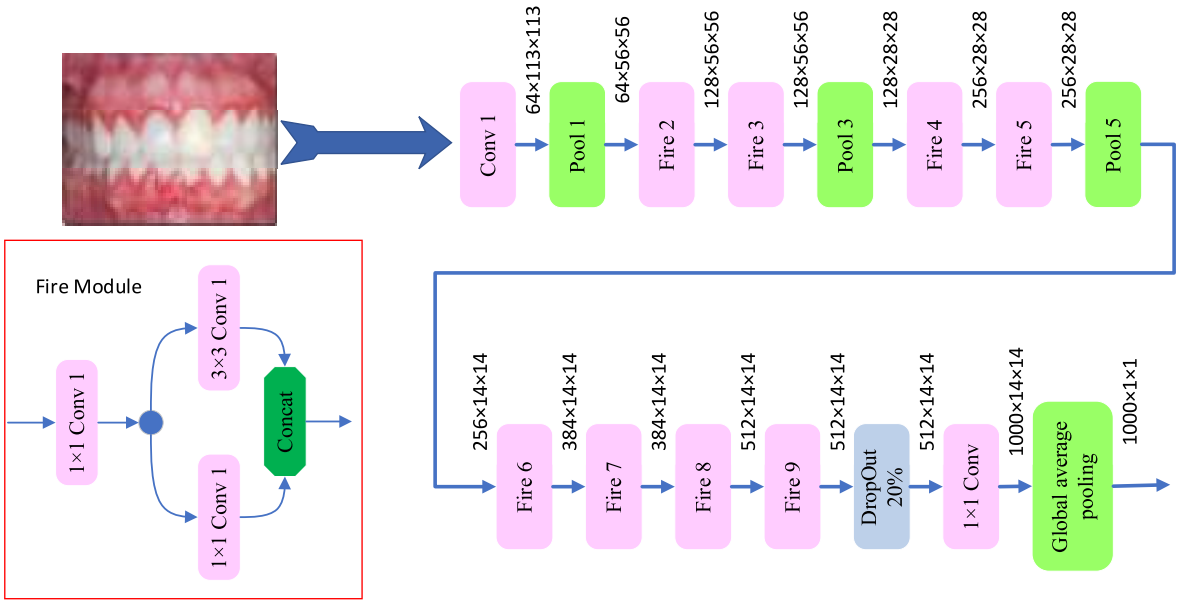


Fig. 9. Architecture of the SqueezeNet model.

trainable parameters in DenseNet is fewer than in CNNs. These factors enable the DenseNet architecture to effectively extract features from images and achieve high accuracy in image classification. There are various versions of DenseNet, each with a different number of layers. DenseNet169 is one such version, comprising 169 layers, and has demonstrated good performance in detection. In this model, the fully connected layers are removed by selecting the 'include_top = False' argument. Instead, a Dense512 layer is added, followed by a ReLU layer and a Softmax layer. The architecture diagram of the DenseNet169 model is depicted in Fig. 8.

SqueezeNet: The SqueezeNet architecture, introduced in 2016 by Forrest Iandola and his team at Stanford University [42], was designed to significantly reduce the model parameters without compromising accuracy. SqueezeNet comprises densification modules known as Fire, which include a 1x1 densification layer and a combination of 3x3 and 1x1 layers. These modules densify the information to reduce the input channels and then expand the densified information to extract features. By executing compression and reducing the number of parameters, SqueezeNet maintains good quality in feature detection. This model is particularly suitable for applications that demand high speed and less volume. In this model, a custom DropOut layer of 20 % is incorporated to prevent overfitting. The architecture diagram of the SqueezeNet model is depicted in Fig. 9.

3.3. Proposed NFR-EDL model

In this research, an ensemble method for image classification is proposed with the primary objective of enhancing classification accuracy by merging the results of several distinct models. This method employs four prevalent CNN models, namely VGG16, DenseNet169, ResNet50, and SqueezeNet. All layers of these base classifiers, except for the last five, are pre-trained on the ImageNet dataset. The final five layers of the base learners are fine-tuned on the datasets. Given that the original dataset is relatively small and has been expanded using data augmentation, a 5-fold cross-validation equivalence is utilized.

In the proposed approach, a combined system based on fuzzy logic is employed to merge the outputs of different models. Each model generates a classification probability for each class. Initially, the output of each model comprises classification probabilities for different classes. These probabilities are converted into fuzzy numbers through three non-linear functions, namely the decaying function, the tanH function, and the Sigmoid function. These functions chosen for constructing rank scores were selected based on several key factors: they capture different properties of confidence scores, effectively handling various distributions; map inputs to a bounded range for fair comparison of base models; emphasize high-confidence predictions while reducing the impact of low-confidence ones; support gradient-based optimization for fine-tuning ensemble weights; have proven effective in machine learning, particularly in confidence calibration and ensemble methods; and are computationally efficient with minimal overhead.

The resulting fuzzy numbers represent the degree of confidence or rank of each model for each class. Additionally, this non-linear conversion reduces the variability of the model outputs and enhances confidence in the results [43–45]. The three non-linear functions used in the proposed architecture are illustrated in Eqs. (1)–(3). These three functions compute the total deviation from the expected value. A lesser deviation indicates a higher confidence in a specific class. A mapped value indicates proximity to 1; another value indicates deviation from 1. Subsequently, the fuzzy scores related to each class for all models are summed up, and the class with the lowest total score is selected as the final output. With this method, the prediction errors of individual models are minimized, and classification accuracy is increased. Overall, the proposed fuzzy ranking system, utilizing non-linear functions and an intelligent combination of model outputs, provides superior performance compared to traditional ranking methods. For a better understanding, the points are explained as follows:

In our NFR-EDL model, the confidence scores represent the probability distribution output by each base model for a given input image, spanning all possible classes. These scores are derived using the Softmax function, which is applied to the final layer output of each CNN base model. Specifically, X_1^m, X_2^m, X_c^m, and X_C^m, denote the confidence scores predicted by the m-th base model for each class c, where $m = 1,2,3,4$ represents the four base models and $c = 1, 2, \ldots, C$ corresponds to the total number of classes. For each input image i, $\left(X_c^m\right)^i = softmax\left(Z_c^m\right)^i$, where Z_c^m is the final layer output of the m-th CNN base model for c-th class.

Initially, all the confidence scores of the individual base learners related to individual images are collected and mapped using three non-linear equations, as follows:

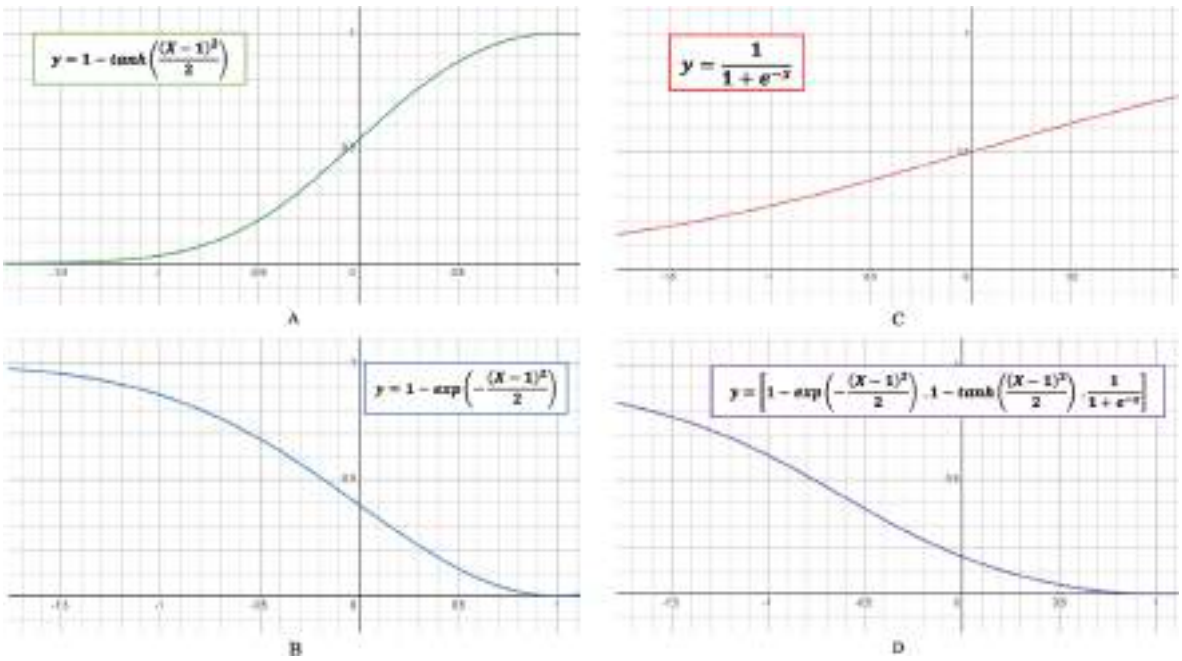


Fig. 10. Non-linear functions used to create fuzzy ranks in the proposed framework. X represents the predicted probability of a class from the sample data and y represents the product of the fuzzy rank.

$$R_k^{i1} = 1 - tanh\left(\frac{\left(X_k^i - 1\right)^2}{2}\right) \tag{1}$$

$$R_k^{i2} = 1 - exp\left(-\frac{\left(X_k^i - 1\right)^2}{2}\right) \tag{2}$$

$$R_k^{i3} = \frac{1}{1 + e^{-X_k^i}} \tag{3}$$

where $\left(R_1^{i_1}, R_2^{i_1}, R_3^{i_1}, \ldots, R_c^{i_1}\right)$, $\left(R_1^{i_2}, R_2^{i_2}, R_3^{i_2}, \ldots, R_c^{i_2}\right)$, and $\left(R_1^{i_3}, R_2^{i_3}, R_3^{i_3}, \ldots, R_c^{i_3}\right)$ are three non-linear fuzzy ranks.

The first non-linear function, given in Eq. (1), is the tanH function. This function is an increasing function, meaning that it increases with the increase in confidence score or probability of belonging to a class. Essentially, it acts as a reward function, where the closer the confidence score, or X, is to 1, the greater the reward, as depicted in Fig. 10-A. The second non-linear function, given in Eq. (2), is a decreasing function that calculates the deviation from 1. In other words, if X is close to 0, the deviation will be greater, as shown in Fig. 10-B. The third non-linear function is a sigmoid function, used to normalize the overall function, as shown in Fig. 10-C. The nonlinear functions chosen for constructing rank scores in our NFR-EDL method were selected based on several key factors: they capture different properties of confidence scores, effectively handling various distributions; map inputs to a bounded range for fair comparison of base models; emphasize high-confidence predictions while reducing the impact of low-confidence ones; support gradient-based optimization for fine-tuning ensemble weights; have proven effective in machine learning, particularly in confidence calibration and ensemble methods; and are computationally efficient with minimal overhead. Furthermore, $Rank\_Score_k^i$ represents the fused rank scores, which is given as follows:

$$Rank_{Score_k}{}^i = R_1^i \times R_2^i \times R_3^i \tag{4}$$

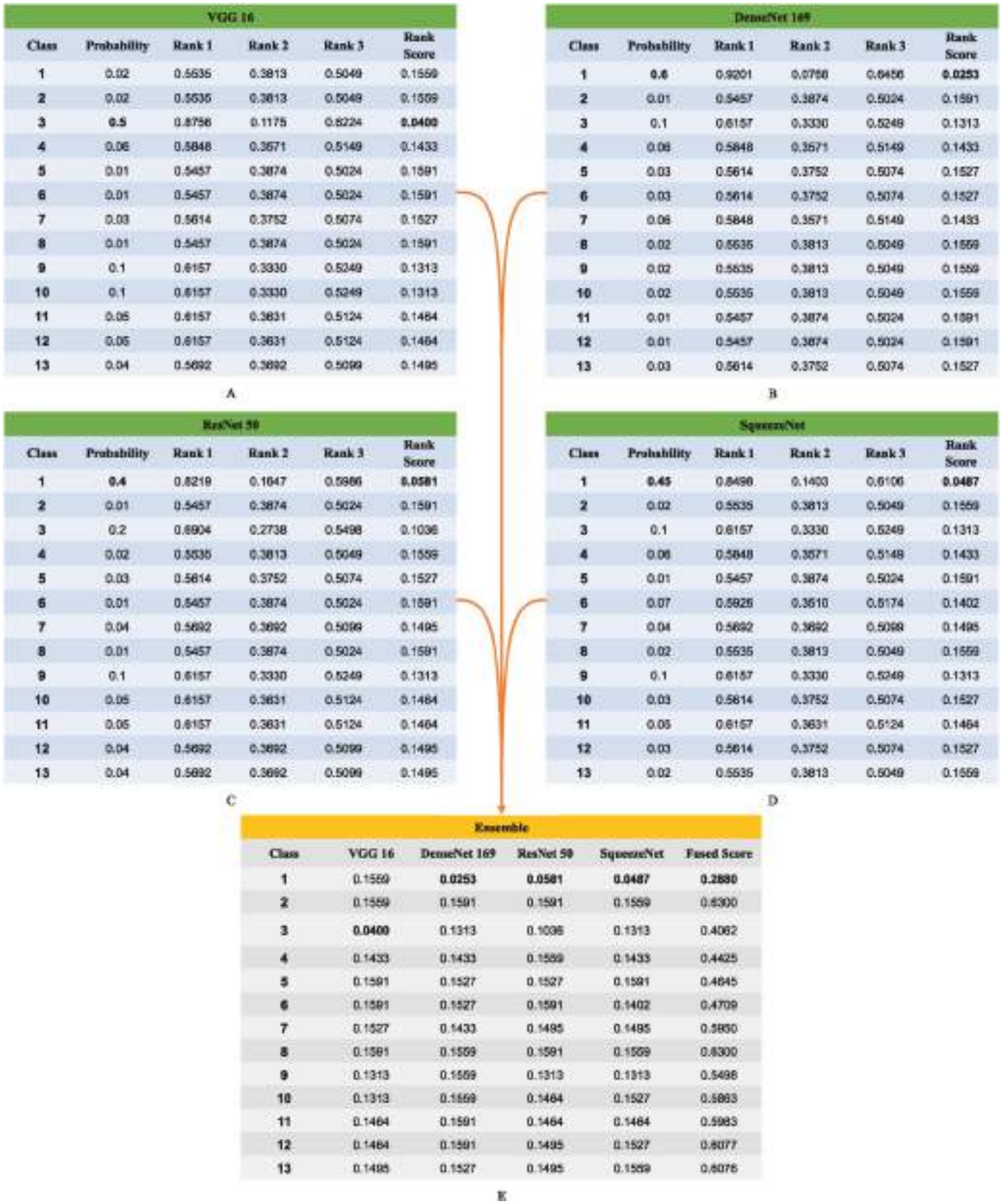
In Eq. (4), the product of three non-linear functions is obtained. This function is a decreasing function and decreases with the increase in probability. The product of the three non-linear functions is also shown in Fig. 10-D. The fused coefficients are (FS_1,FS_2, FS_3,..., FS_C), where FS_k is given as follows:

$$FS_k = \sum_{i=1}^{m} Rank_{Score_k}{}^i, \forall k = 1, 2, 3, \ldots, c \tag{5}$$

This fused score can be used as the final score for each class. Then, the class with the lowest fused score is considered the winner using Eq. (6).

$$Class(I) = \min[FS_k], \forall k = 1, 2, 3, \ldots, c \tag{6}$$

The algorithm is illustrated in Fig. 11. First, the probability scores are gathered from the four base models. Equations (1)–(3) are utilized to compute rank-1, rank-2, and rank-3, respectively. Similarly, the rank score for all base models of each class is computed using Eq. (4). Subsequently, the fused Score for each class is calculated using Eq. (5). The final winner is identified based on the fused score, as outlined in Eq. (6). Finally, the pseudo-code for calculating the fuzzy rank is provided in



VGG 16

Class	Probability	Rank 1	Rank 2	Rank 3	Rank Score
1	0.02	0.5635	0.3813	0.5049	0.1559
2	0.02	0.5635	0.3813	0.5049	0.1559
3	**0.5**	0.8756	0.1175	0.8224	**0.0400**
4	0.06	0.5848	0.3571	0.5149	0.1433
5	0.01	0.5457	0.3874	0.5024	0.1591
6	0.01	0.5457	0.3874	0.5024	0.1591
7	0.03	0.5614	0.3752	0.5074	0.1527
8	0.01	0.5457	0.3874	0.5024	0.1591
9	0.1	0.6157	0.3330	0.5249	0.1313
10	0.1	0.6157	0.3330	0.5249	0.1313
11	0.05	0.6157	0.3631	0.5124	0.1464
12	0.05	0.6157	0.3631	0.5124	0.1464
13	0.04	0.5692	0.3692	0.5099	0.1495

A

DenseNet 169

Class	Probability	Rank 1	Rank 2	Rank 3	Rank Score
1	**0.6**	0.9201	0.0758	0.8456	**0.0253**
2	0.01	0.5457	0.3874	0.5024	0.1591
3	0.1	0.6157	0.3330	0.5249	0.1313
4	0.06	0.5848	0.3571	0.5149	0.1433
5	0.03	0.5614	0.3752	0.5074	0.1527
6	0.03	0.5614	0.3752	0.5074	0.1527
7	0.06	0.5848	0.3571	0.5149	0.1433
8	0.02	0.5635	0.3813	0.5049	0.1559
9	0.02	0.5635	0.3813	0.5049	0.1559
10	0.02	0.5635	0.3813	0.5049	0.1559
11	0.01	0.5457	0.3874	0.5024	0.1591
12	0.01	0.5457	0.3874	0.5024	0.1591
13	0.03	0.5614	0.3752	0.5074	0.1527

B

ResNet 50

Class	Probability	Rank 1	Rank 2	Rank 3	Rank Score
1	**0.4**	0.8219	0.1647	0.5986	**0.0581**
2	0.01	0.5457	0.3874	0.5024	0.1591
3	0.2	0.6904	0.2738	0.5498	0.1036
4	0.02	0.5635	0.3813	0.5049	0.1559
5	0.03	0.5614	0.3752	0.5074	0.1527
6	0.01	0.5457	0.3874	0.5024	0.1591
7	0.04	0.5692	0.3692	0.5099	0.1495
8	0.01	0.5457	0.3874	0.5024	0.1591
9	0.1	0.6157	0.3330	0.5249	0.1313
10	0.05	0.6157	0.3631	0.5124	0.1464
11	0.05	0.6157	0.3631	0.5124	0.1464
12	0.04	0.5692	0.3692	0.5099	0.1495
13	0.04	0.5692	0.3692	0.5099	0.1495

C

SqueezeNet

Class	Probability	Rank 1	Rank 2	Rank 3	Rank Score
1	**0.45**	0.8498	0.1403	0.6106	**0.0487**
2	0.02	0.5635	0.3813	0.5049	0.1559
3	0.1	0.6157	0.3330	0.5249	0.1313
4	0.06	0.5848	0.3571	0.5149	0.1433
5	0.01	0.5457	0.3874	0.5024	0.1591
6	0.07	0.5826	0.3610	0.5174	0.1402
7	0.04	0.5692	0.3692	0.5099	0.1495
8	0.02	0.5635	0.3813	0.5049	0.1559
9	0.1	0.6157	0.3330	0.5249	0.1313
10	0.03	0.5614	0.3752	0.5074	0.1527
11	0.05	0.6157	0.3631	0.5124	0.1464
12	0.03	0.5614	0.3752	0.5074	0.1527
13	0.02	0.5635	0.3813	0.5049	0.1559

D

Ensemble

Class	VGG 16	DenseNet 169	ResNet 50	SqueezeNet	Fused Score
1	0.1559	**0.0253**	**0.0581**	**0.0487**	**0.2880**
2	0.1559	0.1591	0.1591	0.1559	0.6300
3	**0.0400**	0.1313	0.1036	0.1313	0.4062
4	0.1433	0.1433	0.1559	0.1433	0.4425
5	0.1591	0.1527	0.1527	0.1591	0.4645
6	0.1591	0.1527	0.1591	0.1402	0.4709
7	0.1527	0.1433	0.1495	0.1495	0.5950
8	0.1591	0.1559	0.1591	0.1559	0.6300
9	0.1313	0.1559	0.1313	0.1313	0.5498
10	0.1313	0.1559	0.1464	0.1527	0.5863
11	0.1464	0.1591	0.1464	0.1464	0.5983
12	0.1464	0.1591	0.1495	0.1527	0.6077
13	0.1495	0.1527	0.1495	0.1559	0.6076

E

Fig. 11. Calculation of rank scores from four base models and subsequently the overall fused score for making the final decision about the classification result. The winner constructed by DenseNet169, ResNet50, and SqueezeNet is class 1, but VGG16 chooses class 3. The proposed NFR-EDL method works correctly by choosing class 1 as the winner.

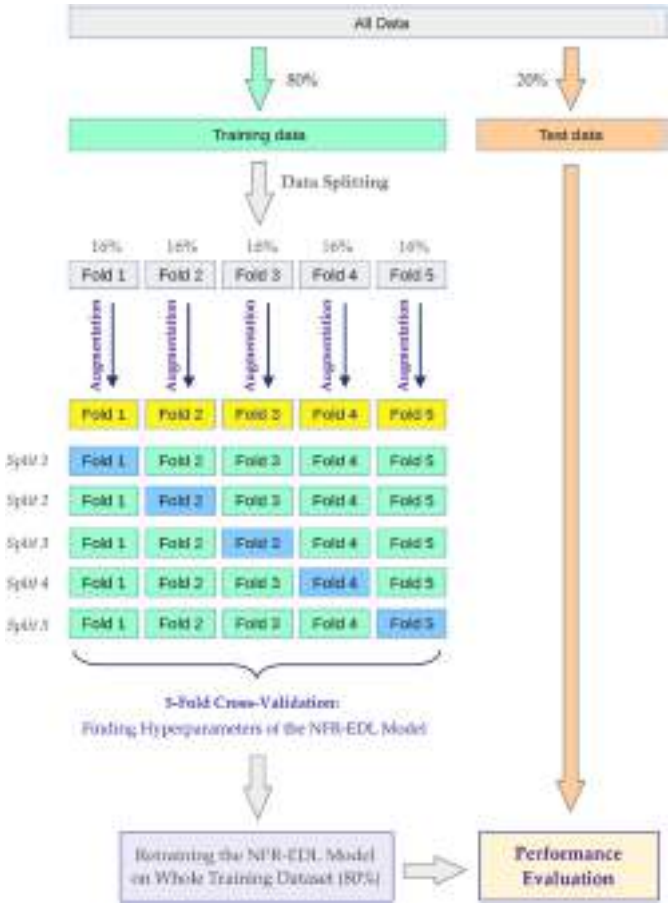


Fig. 12. The data splitting and 5-fold cross-validation method used in the NFR-EDL model.

Algorithm 1.

Algorithm 1. Calculating Fuzzy rank with 5-fold cross-validation

This algorithm calculates the fuzzy rank for each class based on the confidence scores from multiple base models, using 5-fold cross-validation to improve the robustness of the predictions

Number of base models: m,

Number of classes: c,

Confidence scores: $(X_1^m, X_2^m, X_3^m, \ldots, X_c^m)$

Number of folds: n

Set: Rank_Score ← Empty_list

Fused_Score ← Empty_list for each fold i from 1 to n do

for each base model j from 1 to m do

for each class k from 1 to c do

Calculate $R_k^{i1} = 1 - tanh\left(\frac{(X_k^i - 1)^2}{2}\right)$

Calculate $R_k^{i2} = 1 - exp\left(-\frac{(X_k^i - 1)^2}{2}\right)$

Calculate $R_k^{i3} = \frac{1}{1 + e^{-X_k^i}}$

Calculate $Rank\_Score_k^i = R_1^i \times R_2^i \times R_3^i$ end

end

for each class k from 1 to c do

Calculate Fused_Score_K $= \sum_{i=1}^{m} Rank\_Score_k^i, \forall k = 1, 2, 3, \ldots, c.$ end

Find the class with the minimum *Fused_Score_K*, and assign it as class(I) for fold i end

3.4. Experimental setting

This section outlines the setup used to implement the proposed NFR-EDL model for detecting oral and dental diseases. As shown in Fig. 12, we split the dataset into 80 % for training and 20 % for testing to evaluate the performance of the NFR-EDL model. The training set was further divided into five equal parts, with each fold representing 16 % of the entire dataset. We made sure that each fold had a balanced mix of samples from all classes, ensuring that no category was over- or under-represented. To make the model more robust, we applied data augmentation techniques to each fold individually within the training set. This not only enriches the dataset but also prevents any overlap or leakage of information between the training, validation, and testing sets. Importantly, to avoid data leakage and ensure fairness, we grouped all images from the same patient together and assigned them to a single fold. This precaution guarantees that no patient's data appears in multiple folds, preserving the integrity of the evaluation process. During the cross-validation process, four folds are used for training the model and one remaining fold for validation. After tuning the model's hyperparameters with this setup, we retrained the model on the entire 80 % training set and evaluated its performance on the reserved 20 % unseen test set.

Initially, the data is divided into 5-folds using cross-validation. Four base models are then employed to perform the classification of oral and dental diseases. The Adam optimizer, with an initial learning rate of 0.0001, is used to minimize the loss function. An early stopping strategy is applied, and each model is trained for a maximum of 50 epochs with a batch size of 16. The Cross-Entropy Loss is minimized for the Loss Function with a learning rate of 0.0001. The hyperparameter values employed for training the base CNN classifiers are presented in Table 2.

Table 2
Hyperparameter values used for training the base CNN classifiers.

Hyperparameter	Value
Optimizer	Adam
Learning rate	0.0001
Loss Function	Cross Entropy Loss
Number of epochs	50
Batch size	16
DropOut	20 %

3.5. Performance measures

The proposed approach to validate the information involved comparing the anticipated labels to the context labels, which resulted in one of four outcomes.

- True Positive (TP): The number of positive examples correctly identified by the model.
- False Negative (FN): The number of examples that the model mistakenly classified as negative.
- False Positive (FP): The number of examples that the model mistakenly classified as positive.
- True Negative (TN): The number of examples that the model correctly identified as negative.

The model's accuracy can be calculated using Eq. (7), which divides the number of total predictions produced by the number of right predictions.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

The precision is the ratio of true positives to all positive forecasts, Eq. (8) shows this equals the number of true positives divided by the sum of true positives plus false positives.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

Recall is the quantitative way in which a model can predict all of the positive cases in the dataset. The model's number of genuine positive predictions divided by the total number of positive cases within the

Table 3
Results of the different techniques in different validation folds of the 5-fold cross-validation method in Kaggle.

Model	Metric	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	(mean ± std)Average
VGG16	Accuracy %	93.93	94.45	94.45	94.27	94.36	(94.29 ± 0.22)
	Precision %	93.52	94.60	94.16	94.38	94.20	(94.17 ± 0.40)
	Recall %	94.23	94.34	94.68	94.12	94.46	(94.37 ± 0.21)
	F1 Score %	93.92	94.42	94.40	94.25	94.37	(94.27 ± 0.20)
DenseNet169	Accuracy %	96.13	96.45	95.70	96.13	95.70	(96.02 ± 0.32)
	Precision %	96.42	97.18	95.94	96.27	95.85	(96.33 ± 0.52)
	Recall %	95.73	96.04	95.50	96.03	95.70	(95.80 ± 0.23)
	F1 Score %	96.10	96.53	95.68	96.14	95.75	(96.04 ± 0.34)
ResNet50	Accuracy %	95.49	96.82	95.53	95.53	95.53	(95.78 ± 0.58)
	Precision %	95.24	96.95	95.35	95.24	95.64	(95.68 ± 0.72)
	Recall %	95.68	96.62	95.83	95.73	95.43	(95.85 ± 0.45)
	F1 Score %	95.45	96.80	95.52	95.52	95.53	(95.76 ± 0.58)
SqueezeNet	Accuracy %	92.18	92.61	91.71	91.83	92.09	(92.08 ± 0.35)
	Precision %	92.03	92.65	91.74	92.71	92.54	(92.33 ± 0.42)
	Recall %	92.28	92.51	91.60	92.92	92.00	(92.26 ± 0.50)
	F1 Score %	92.17	92.63	91.70	92.81	92.19	(92.30 ± 0.43)
NFR-EDL (Proposed)	Accuracy %	97.83	98.09	97.70	97.22	97.06	(97.57 ± 0.43)
	Precision %	97.61	97.93	97.82	97.03	96.84	(97.45 ± 0.48)
	Recall %	97.93	98.29	97.60	97.42	97.26	(97.70 ± 0.41)
	F1 Score %	97.85	98.07	97.68	97.21	97.09	(97.58 ± 0.42)

dataset defines the recall. This is calculated by dividing the TP by the sum of the True Positive and the FN values. The recall is calculated as follows:

$$Recall = \frac{TP}{FN + TP} \tag{9}$$

The F1 Score is commonly defined as the harmonic mean of Precision and Recall values. The F1 Score is calculated using Eq. (10).

$$F1\ Score = \frac{2{*}Precision{*}recall}{Precision + recall} \tag{10}$$

The Receiver Operating Characteristic (ROC) curve is a critical tool for evaluating the performance of classification models, showing how changes in decision thresholds affect the True Positive Rate (TPR) and False Positive Rate (FPR). Furthermore, the AUC value is used to describe the predictive capabilities of the model, which ranges from 0 to 1, and values close to 1 indicate high model ability. This paper calculates the Micro-averaged AUC value that provides the combined performance of the model across all classes and provides a better understanding of the overall performance of the model.

4. Performance evaluation

4.1. Results on Kaggle

The results in Table 3 provide a detailed assessment of the proposed NFR-EDL model compared to its base models without the fuzzy branch (VGG16, DenseNet169, ResNet50, and SqueezeNet) using the 5-fold cross-validation. These results underline the architecture's efficiency in feature extraction and classification during the validation phase. Furthermore, Table 3 showcases the performance metrics for each model across five folds. Among the base models, DenseNet169 stands out as the most competitive, achieving an average accuracy of 96.02 % and an F1 Score of 96.04 %. ResNet50 follows closely with an average accuracy of 95.78 %, though its higher standard deviations indicate less stability compared to DenseNet169. In contrast, the proposed NFR-EDL model consistently outperforms all base models, achieving an average accuracy of 97.57 % with minimal variability across folds. Its precision (97.45 %), recall (97.70 %), and F1 Score (97.58 %) further highlight its balanced and robust performance, excelling at minimizing both false positives and false negatives. The low standard deviations across metrics underscore the model's resilience to variations in data subsets, reflecting its ability to generalize effectively during training and validation.

After the training and validation phase, the NFR-EDL model's performance on unseen data is assessed by training it on the augmented training set and testing it on the 20 % held-out test set (as illustrated in Fig. 12). For the training process, Fig. 13 provides insights into the learning dynamics of the different base models that contribute to NFR-EDL's design. The results show that DenseNet169 and ResNet50 exhibit less gaps between training and validation accuracies than the VGG16 and SqueezeNet.

The test results presented in Figs. 14–16 provide the outcomes of the unseen test data samples, validating the NFR-EDL's generalization capabilities. The confusion matrices in Fig. 14 provide a detailed view of how each model performs on individual classes, showing that NFR-EDL consistently achieved the lowest misclassification rates across all categories. This suggests that the proposed model excels at handling the complex patterns in the dataset, minimizing errors when distinguishing between similar dental conditions. In Fig. 15, the comparison of ROC curves further highlights the NFR-EDL's superior performance, with a steeper curve and a higher AUC compared to the base models, indicating its stronger ability to differentiate between classes. The per-class ROC analysis for NFR-EDL also shows consistently high AUC values across all categories, underscoring its balanced and reliable performance in detecting a wide range of dental conditions. Finally, Fig. 16 summarizes the overall test set performance, where NFR-EDL stands out with higher metrics across the board, clearly outperforming the base models. With an accuracy of 97.08 %, NFR-EDL again outperformed all base models, achieving the highest precision (97.11 %), recall (97.08 %), and F1 Score (97.07 %). Among the base models, ResNet50 performed closely behind with an accuracy of 96.82 %, while DenseNet169 followed with 96.35 %. Furthermore, VGG16 and SqueezeNet showed lower performance, achieving accuracies of 95.45 % and 92.61 %, respectively.

These results collectively validate the superiority of the proposed NFR-EDL model, not only during training and validation but also in its ability to generalize effectively to unseen data. This establishes NFR-EDL as a highly reliable and efficient tool for real-world dental disease diagnosis.

4.2. Results on MOD

The results of the different techniques on the test set for the MOD dataset are provided in Figs. 17–19. Fig. 17A shows the ROC curves for all models, demonstrating that NFR-EDL exhibits a sharper rise and

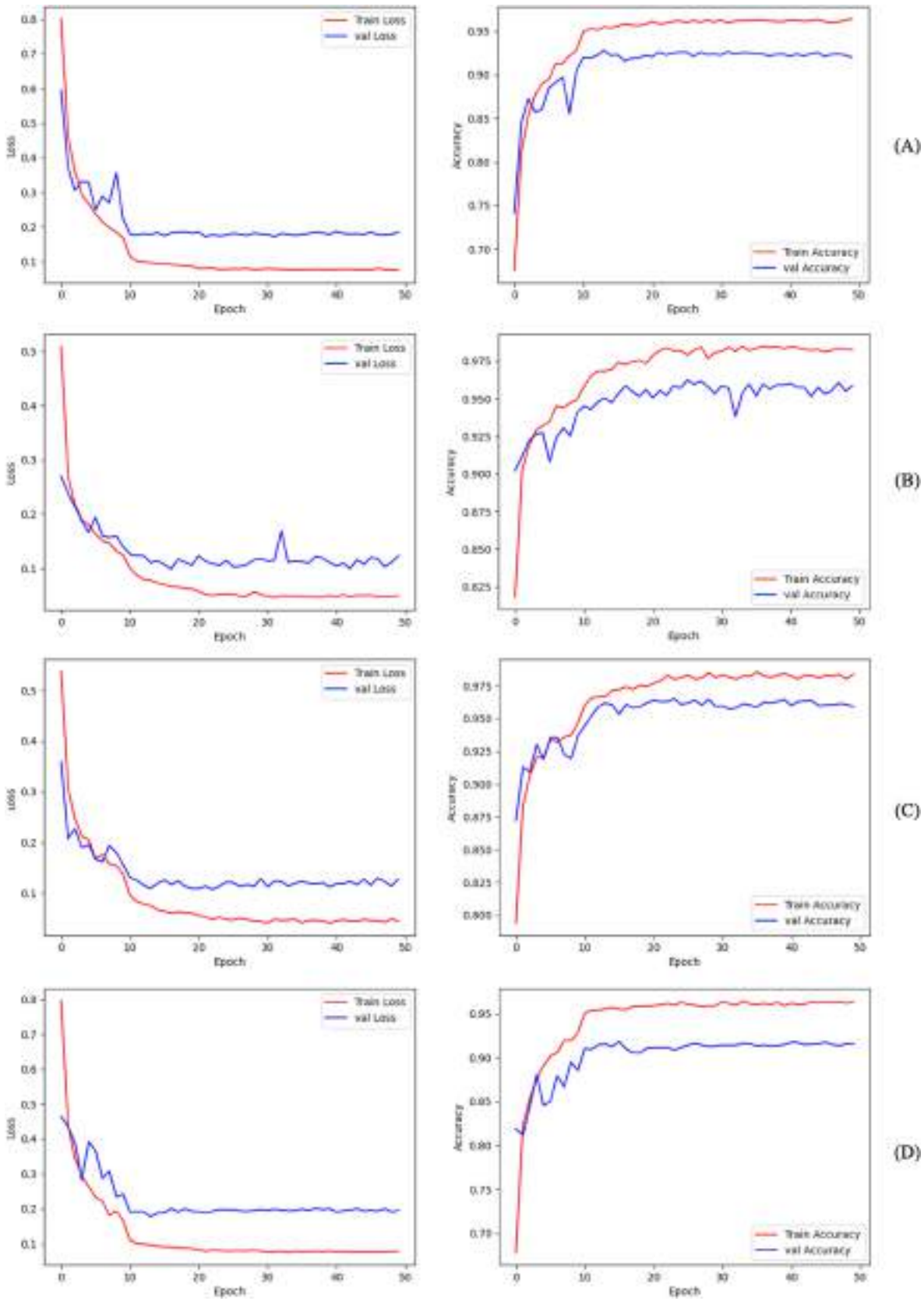


Fig. 13. The loss and accuracy curves for training progress of the four base learners on the whole 80 % training set in Kaggle: A) VGG16, B) DenseNet169, C) ResNet50, and D) SqueezeNet.

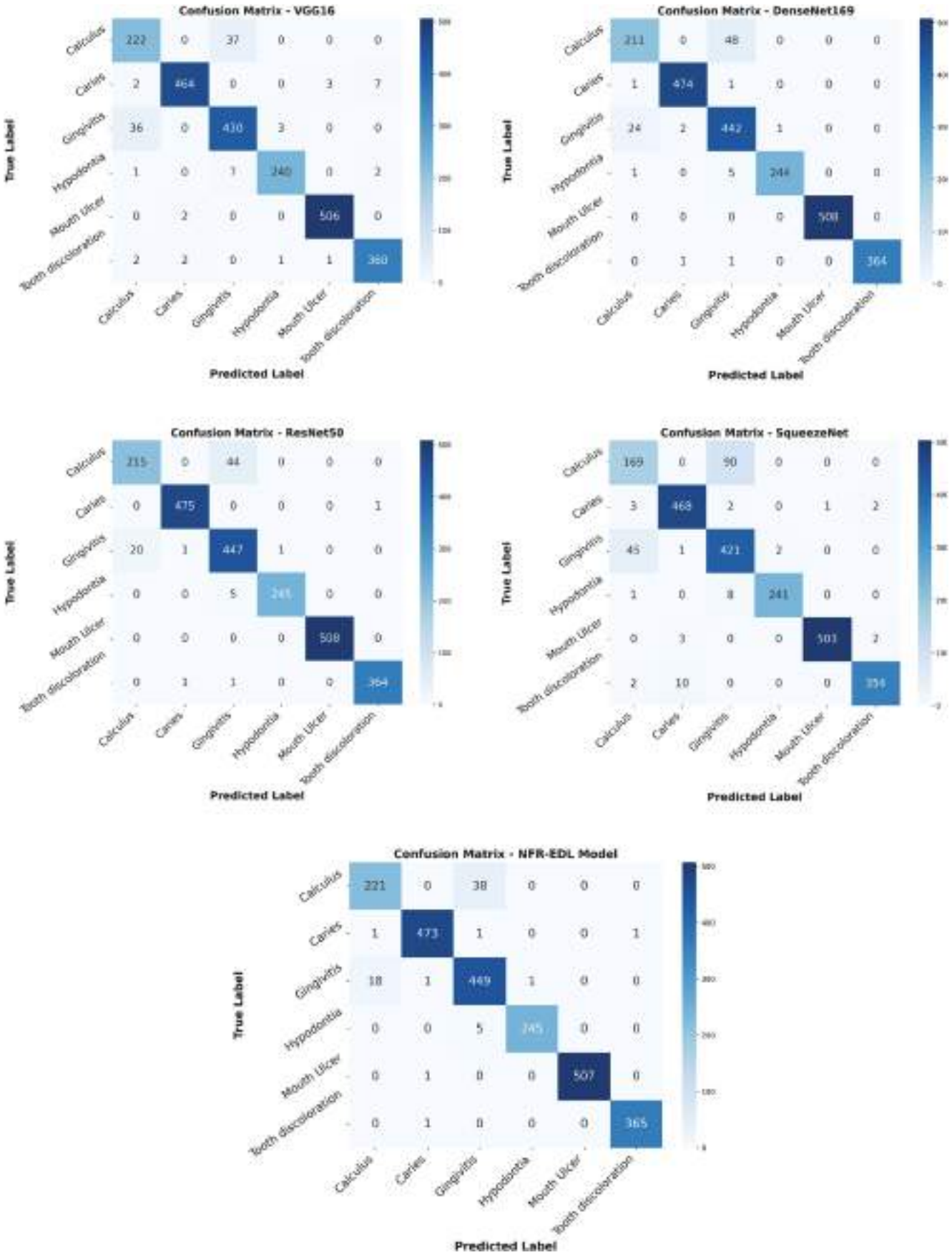


Fig. 14. Confusion matrix of the different techniques on the test set in Kaggle.

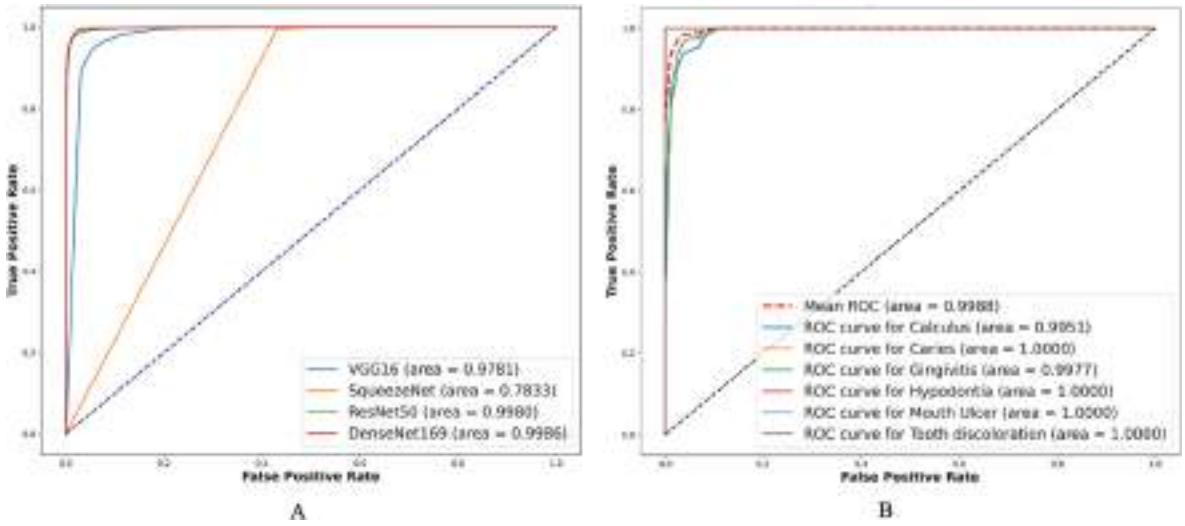
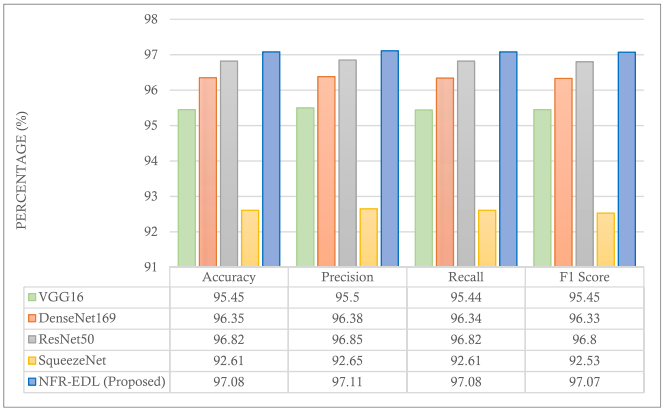
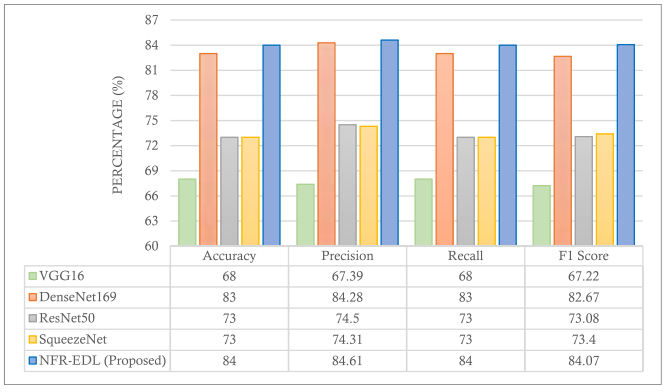


Fig. 15. ROC curves of the different techniques on the test set in Kaggle: A) average ROC curves of the different base learners for all classes, and B) ROC curves of the proposed NFR-EDL model per different classes.



	Accuracy	Precision	Recall	F1 Score
VGG16	95.45	95.5	95.44	95.45
DenseNet169	96.35	96.38	96.34	96.33
ResNet50	96.82	96.85	96.82	96.8
SqueezeNet	92.61	92.65	92.61	92.53
NFR-EDL (Proposed)	97.08	97.11	97.08	97.07

Fig. 16. Comparison of the performance of the different techniques on the test set in Kaggle.



	Accuracy	Precision	Recall	F1 Score
VGG16	68	67.39	68	67.22
DenseNet169	83	84.28	83	82.67
ResNet50	73	74.5	73	73.08
SqueezeNet	73	74.31	73	73.4
NFR-EDL (Proposed)	84	84.61	84	84.07

Fig. 18. Comparison of the performance of the different techniques on the test set in MOD.

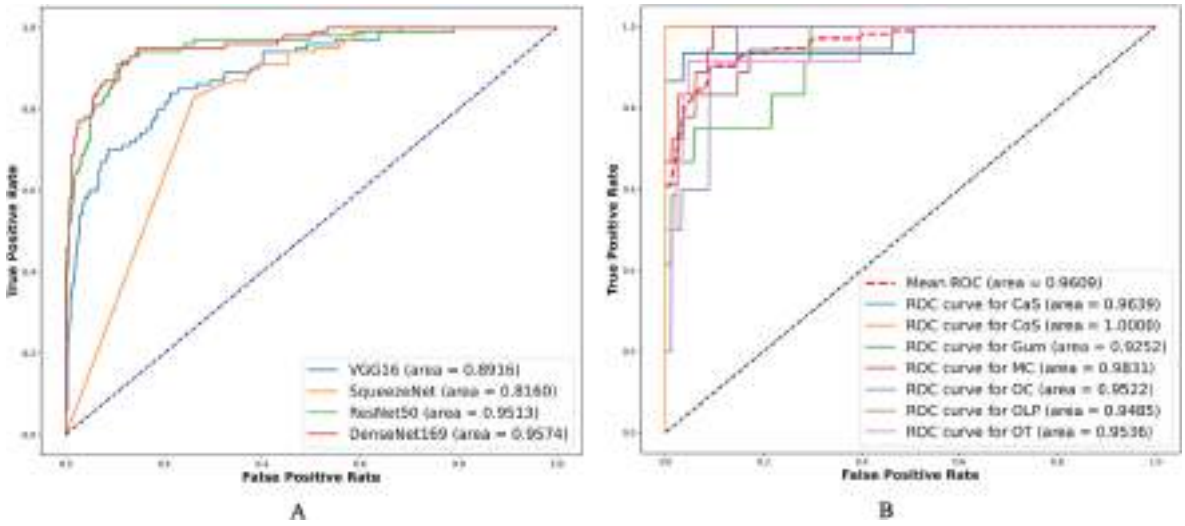


Fig. 17. ROC curves of the different techniques on the test set in MOD: A) average ROC curves of the different base learners for all classes, and B) ROC curves of the proposed NFR-EDL model per different classes.

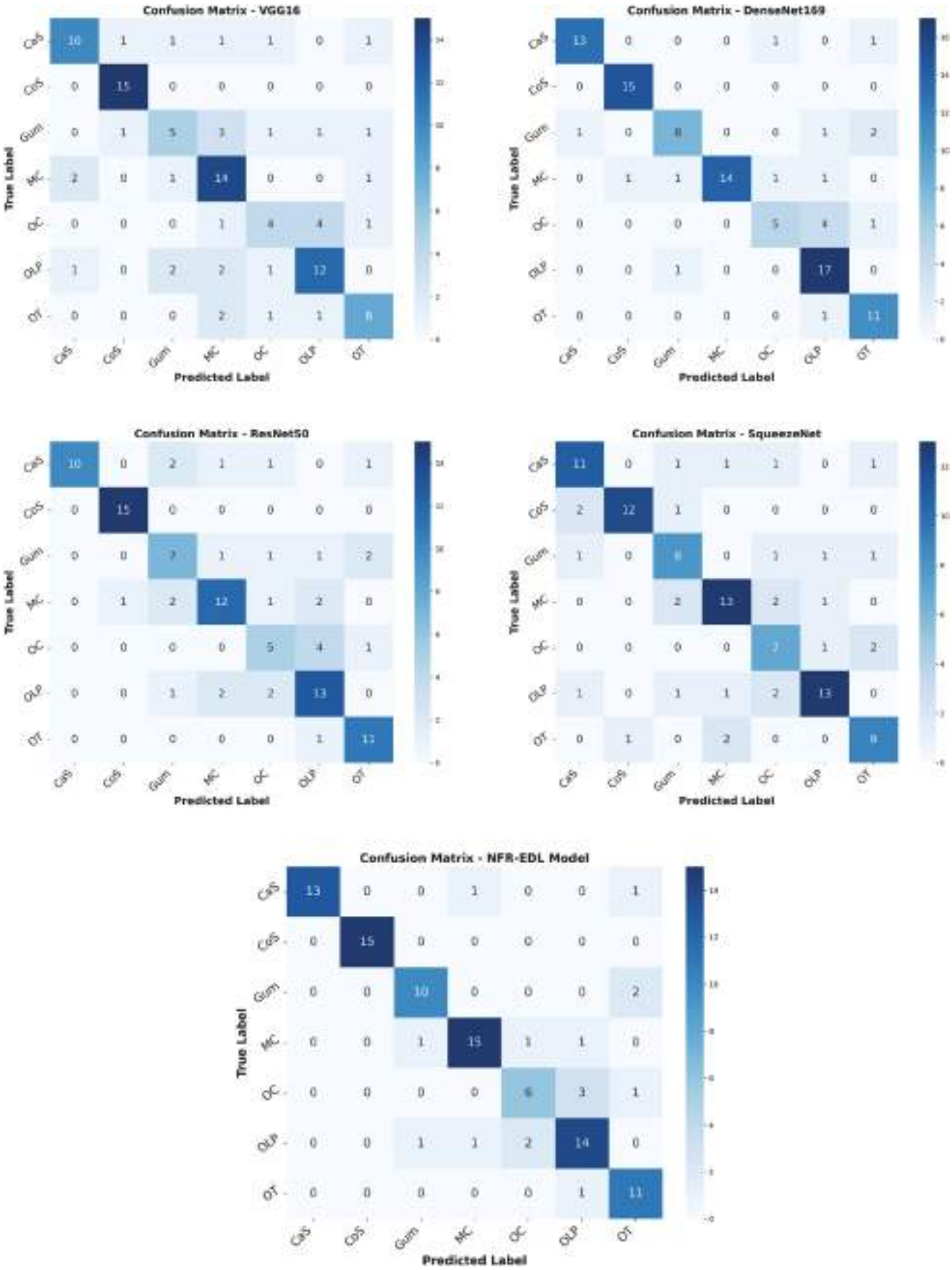


Fig. 19. Confusion matrix of the different techniques on the test set in MOD.

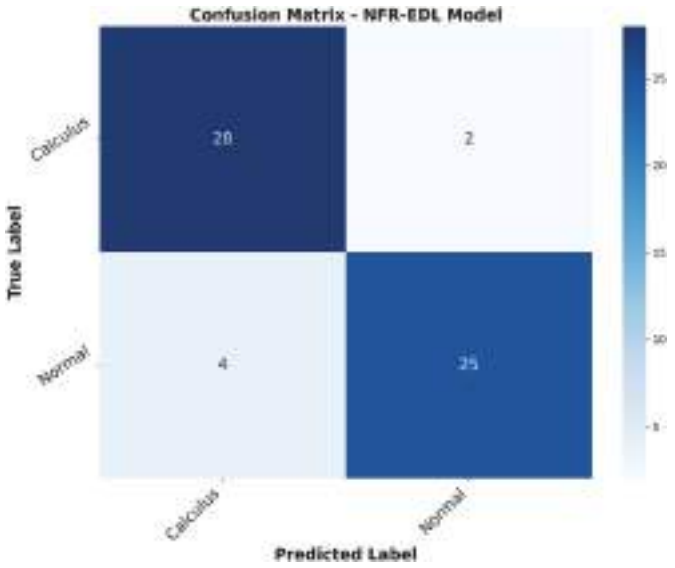
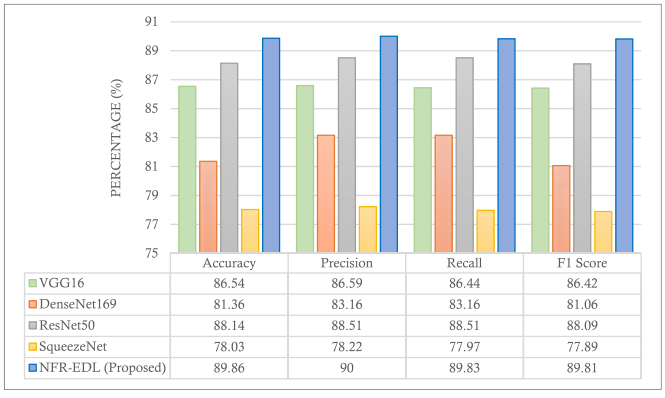


Fig. 20. Confusion matrix of the NFR-EDL model on the test set in ODSI-DB.

greater AUC, indicating its stronger ability to separate classes accurately. Fig. 17B also shows the per-class ROC curves for NFR-EDL, which consistently display high AUC values, reinforcing the model's reliability and robustness across all categories. Furthermore, Fig. 18 presents a visual comparison of the overall performance of the models on the test set, where NFR-EDL's metrics stand out above all others. According to the obtained results, the NFR-EDL model outperforms all the other models in every performance metric. It achieved an accuracy of 84.00 %, precision of 84.61 %, recall of 84.00 %, and an F1 Score of 84.07 %, showing its strong and consistent performance. These results highlight the robustness of NFR-EDL in effectively distinguishing between classes while minimizing both false positives and false negatives. Among the base models, DenseNet169 stood out with the best performance, scoring an F1 Score of 82.67 %, accuracy of 83.00 %, precision of 84.28 %, and recall of 83.00 %. This made it the top competitor contributing to the NFR-EDL model. This comprehensive analysis demonstrates that NFR-EDL generalizes well to unseen test data samples, positioning it as a highly effective solution for dental disease classification on the MOD dataset. The clear distinction in performance between NFR-EDL and the base models underlines its superior ability to manage the variability and nuances within the data, minimizing errors that could arise when classifying similar instances. Looking at the confusion matrix in Fig. 19, we see that the NFR-EDL model achieved the fewest misclassifications across most of the classes, further emphasizing its effectiveness in handling the complexities of the MOD dataset.

4.3. Results on ODSI-DB

The results from the ODSI-DB test set, shown in Figs. 20–22, highlight the strong performance of the NFR-EDL model compared to its base learners. Fig. 20 depicts the confusion matrix of the NFR-EDL model. The ROC curves in Fig. 21 (A and B) provide further confirmation of NFR-EDL's superior performance. In particular, the AUC for NFR-EDL is noticeably higher than that of the base models, which demonstrates the model's ability to differentiate between classes with greater accuracy. The per-class ROC analysis shown in Fig. 21B reinforces this by highlighting that NFR-EDL maintained consistently high AUC scores across both classes. The comparison of overall performance metrics in Fig. 22 shows that the NFR-EDL model outperforms all others in every metric, with an accuracy of 89.86 %, precision of 90.00 %, recall of 89.83 %, and an F1 Score of 89.81 %. ResNet50 came in second, showing competitive results with an accuracy of 88.14 %.



	Accuracy	Precision	Recall	F1 Score
VGG16	86.54	86.59	86.44	86.42
DenseNet169	81.36	83.16	83.16	81.06
ResNet50	88.14	88.51	88.51	88.09
SqueezeNet	78.03	78.22	77.97	77.89
NFR-EDL (Proposed)	89.86	90	89.83	89.81

Fig. 22. Comparison of the performance of the different techniques on the test set in ODSI-DB.

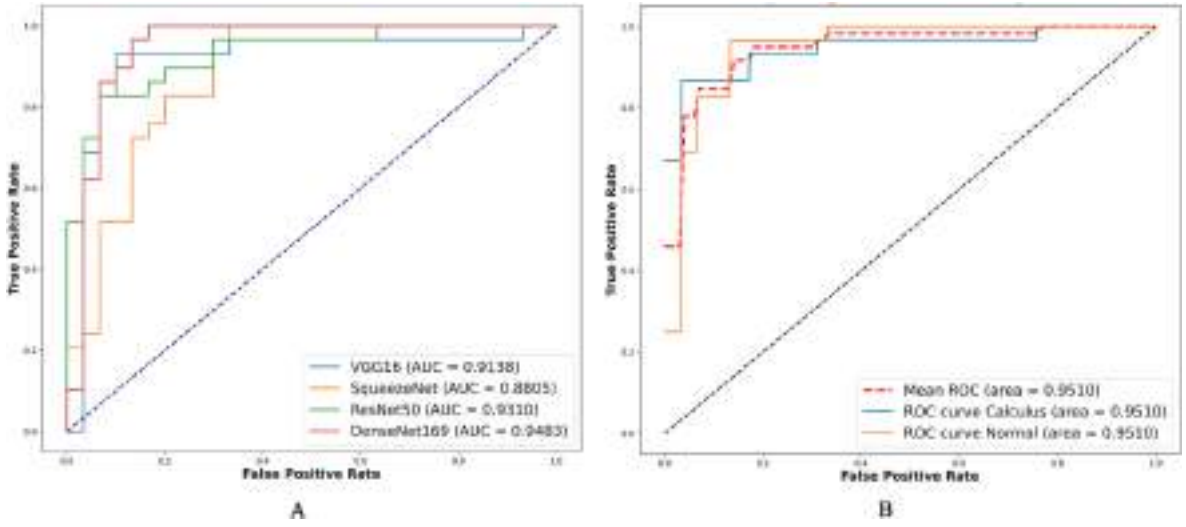


Fig. 21. ROC curves of the different techniques on the test set in ODSI-DB: A) average ROC curves of the different base learners for all classes, and B) ROC curves of the proposed NFR-EDL model per different classes.

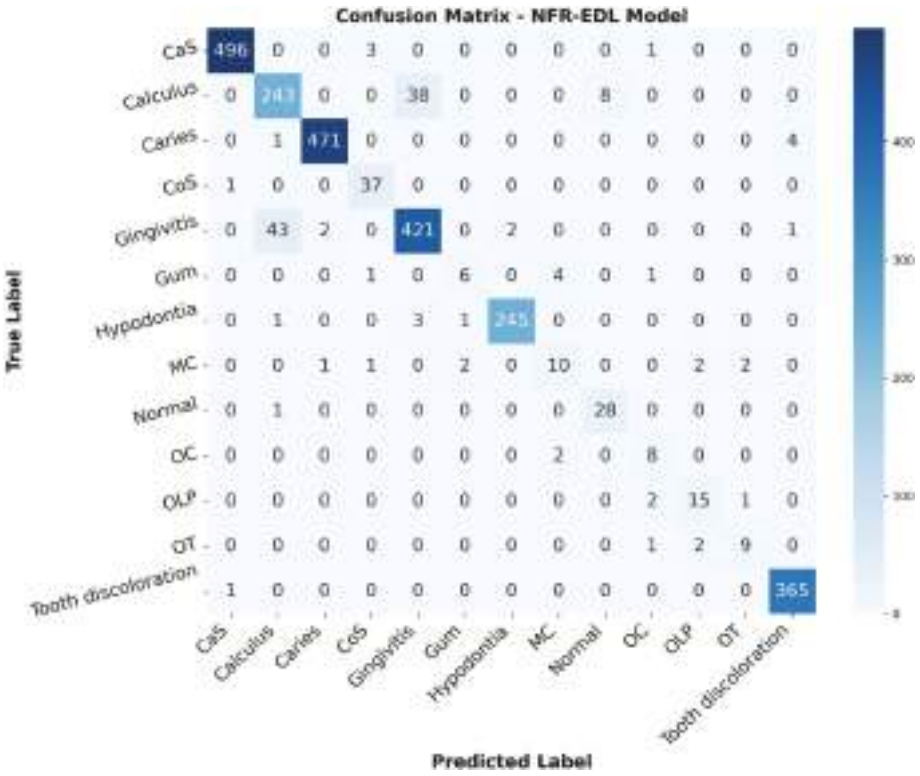


Fig. 23. Confusion matrix of the NFR-EDL model on the test set in OaDD.

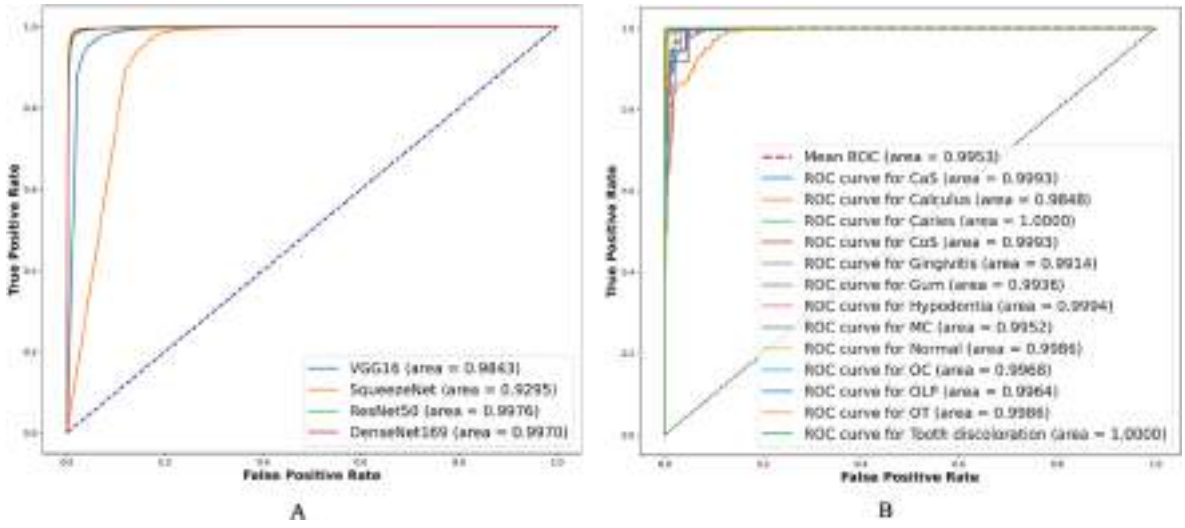


Fig. 24. ROC curves of the different techniques on the test set in OaDD: A) average ROC curves of the different base learners for all classes, and B) ROC curves of the proposed NFR-EDL model per different classes.

4.4. Results on OaDD

The results of the OaDD dataset, which combines data from Kaggle, MOD, and ODSI-DB, are provided in Figs. 23–25. The obtained results further validate the effectiveness of the proposed NFR-EDL model across a broader range of test cases. By outperforming all base models in every evaluation metric, the NFR-EDL model proves itself as a powerful tool for complex oral and dental classification problems. These results suggest that the proposed NFR-EDL model can be generalized effectively to handle datasets that vary in complexity and structure, such as the combined OaDD dataset. This makes NFR-EDL a promising solution for future applications across different fields where high classification accuracy and reliability are paramount.

4.5. Comparison with other ensemble learning techniques

In this section, we compare the performance of the proposed NFR-EDL model against other popular ensemble learning methods, including Random Forest, ADA Boosting, Extra Trees, Ensemble Voting, and Classical Averaging Ensemble. All models are evaluated under the same conditions and applied to the same datasets as the NFR-EDL model to ensure a fair comparison. The obtained results in Table 4 reveal the clear advantages of the NFR-EDL across all metrics on every dataset. Fig. 26, which visually compares the average performance of each model

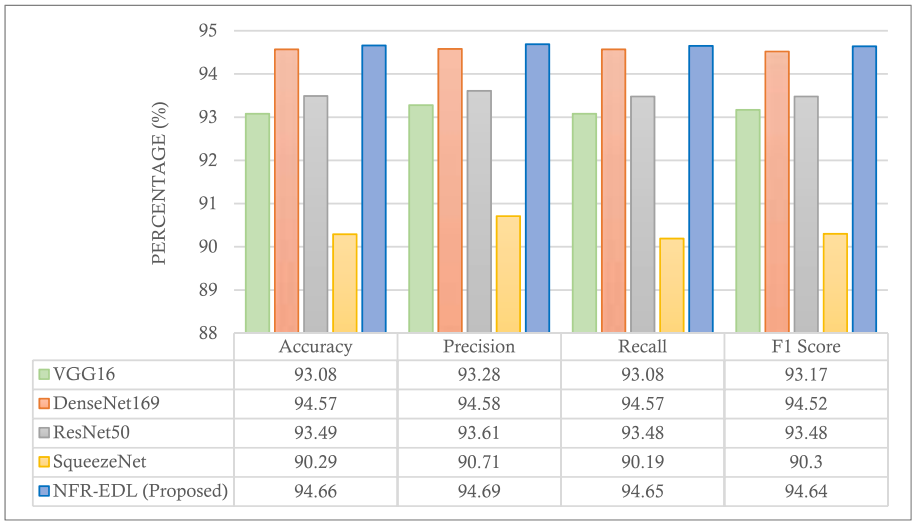


Fig. 25. Comparison of the performance of the different techniques on the test set in OaDD.

Table 4
Comparison of the proposed NFR-EDL model with other ensemble learning techniques.

Model	Metric	Kaggle	MOD	ODSI-DB	OaDD
Random Forest	Accuracy %	94.14	82.00	86.71	91.67
	Precision %	94.24	81.15	87.14	90.51
	Recall %	94.07	83.00	86.17	92.71
	F1 Score %	94.15	82.25	86.67	91.45
ADA Boosting	Accuracy %	94.14	80.40	85.38	91.34
	Precision %	94.96	81.50	84.78	91.02
	Recall %	93.54	79.40	86.35	91.57
	F1 Score %	94.02	80.26	85.88	91.28
Extra Trees	Accuracy %	93.08	79.82	84.32	89.93
	Precision %	93.72	80.50	82.80	88.86
	Recall %	92.58	78.83	86.36	91.53
	F1 Score %	93.16	79.46	84.20	90.10
Ensemble Voting	Accuracy %	93.54	80.32	83.87	91.88
	Precision %	93.24	79.33	84.50	91.24
	Recall %	93.86	81.54	83.05	92.23
	F1 Score %	93.42	80.73	83.70	91.76
Classical Averaging Ensemble	Accuracy %	93.30	81.00	83.05	92.00
	Precision %	93.60	82.51	83.18	92.66
	Recall %	93.30	81.00	83.05	92.05
	F1 Score %	93.45	81.30	83.02	92.28
NFR-EDL (Proposed)	Accuracy %	97.08	84.00	89.86	94.66
	Precision %	97.11	84.61	90.00	94.69
	Recall %	97.08	84.00	89.83	94.65
	F1 Score %	97.07	84.07	89.81	94.64

for all datasets, further corroborates these findings, showing NFR-EDL as the most well-rounded ensemble learning model.

Based on the average performance across all datasets, the proposed NFR-EDL model demonstrates superior accuracy, achieving 91.39 %, which surpasses other ensemble models by notable margins: 2.76 % over

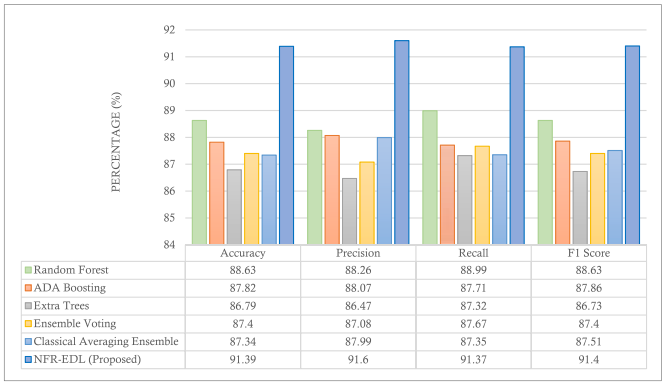


Fig. 26. Comparison of the NFR-EDL model with other ensemble learning techniques (on average for all datasets).

Random Forest (88.63 %), 3.57 % over ADA Boosting (87.82 %), 4.6 % over Extra Trees (86.79 %), 3.99 % over Ensemble Voting (87.4 %), and 4.05 % over Classical Averaging Ensemble (87.34 %). In terms of precision, which indicates the model's ability to correctly identify positive instances, NFR-EDL again leads with an average of 91.6 %, showing improvements of 3.34 %, 3.53 %, 5.13 %, 4.52 %, and 3.61 % over Random Forest, ADA Boosting, Extra Trees, Ensemble Voting, and Classical Averaging Ensemble, respectively. For recall, which reflects the model's ability to detect all relevant instances, NFR-EDL achieves an average of 91.37 %, outperforming the baselines by 2.38 % (Random Forest), 3.66 % (ADA Boosting), 4.05 % (Extra Trees), 3.7 % (Ensemble Voting), and 4.02 % (Classical Averaging Ensemble). The F1 Score, which harmonizes precision and recall, shows a similar trend. NFR-EDL achieves an average of 91.4 %, indicating improvements of 2.77 %, 3.54 %, 4.67 %, 4 %, and 3.89 %, respectively, over the same baseline models. This demonstrates its ability to achieve a strong balance between minimizing false positives and false negatives.

The detailed results presented in Table 4 and visualized in Fig. 26 indicate that NFR-EDL performs consistently across all datasets, striking an optimal balance between precision, recall, and F1 Score. With significantly higher performance across all metrics, NFR-EDL sets a new benchmark for ensemble models, offering a more reliable and accurate solution for complex data recognition in oral and dental diagnostics.

Table 5
Comparison of the proposed NFR-EDL model with existing RGB color photography techniques.

Model	Reference	Metric	Kaggle	MOD	ODSI-DB	OaDD
ResNet34	[30]	Accuracy %	92.83	81.10	84.31	90.11
		Precision %	93.54	81.38	84.44	89.30
		Recall %	91.88	80.67	84.12	90.84
		F1 Score %	92.70	81.04	84.24	89.99
CariesFG	[32]	Accuracy %	97.22	81.65	86.05	92.05
		Precision %	97.67	82.11	86.10	91.66
		Recall %	96.16	80.69	85.99	92.45
		F1 Score %	96.52	81.44	86.07	91.94
OralNet	[34]	Accuracy %	92.80	81.82	85.05	90.84
		Precision %	92.18	80.67	85.43	90.53
		Recall %	93.36	82.27	84.52	91.11
		F1 Score %	92.72	81.50	84.97	90.70
NFR-EDL	Proposed Method	Accuracy %	97.08	84.00	89.86	94.66
		Precision %	97.11	84.61	90.00	94.69
		Recall %	97.08	84.00	89.83	94.65
		F1 Score %	97.07	84.07	89.81	94.64

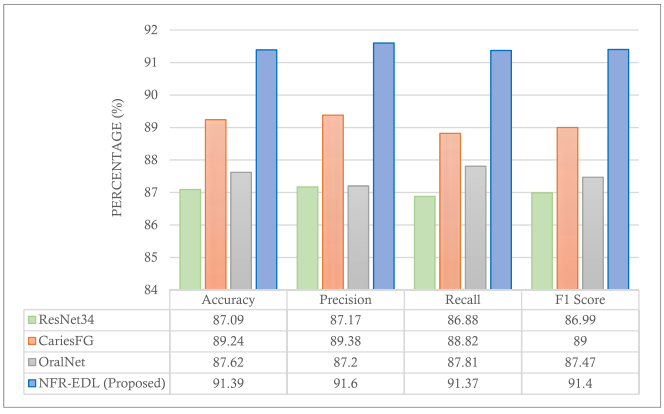


Fig. 27. Comparison of the NFR-EDL model with existing RGB color photography techniques (on average for all datasets). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

4.6. Comparison with existing RGB color photography techniques

In this section, we present a detailed analysis of the performance of the proposed NFR-EDL model by comparing it with three existing deep learning-based techniques that utilize RGB color photography: ResNet34 (Garg et al., 2023) [30], CariesFG (Deepak and Krishna Bhat, 2023) [32], and OralNet (Sampath et al., 2024) [34]. These comparisons are conducted across four datasets under identical experimental conditions, as shown in Table 5. The results indicate that the NFR-EDL model outperforms the other methods in most metrics across different datasets. Although the CariesFG model achieves better accuracy (97.22 %) and precision (97.67 %) on the Kaggle dataset, the NFR-EDL model stands out with its superior balance between precision and recall. This is particularly evident when comparing the F1 Scores of both models on the Kaggle dataset. The NFR-EDL model's ability to maintain a balanced trade-off between precision and recall is crucial in disease diagnosis tasks, where both false positives and false negatives must be minimized to ensure reliable results. One of the key advantages of the NFR-EDL model lies in its consistent performance across all datasets. It excels in F1 Score, indicating that it is not only accurate but also highly reliable in distinguishing between different types of oral and dental diseases.

As illustrated in Fig. 27, the NFR-EDL model outperforms the other models on average across all datasets. The results show that NFR-EDL achieves an average accuracy of 91.39 %, which is 4.30 % higher than ResNet34, 2.15 % higher than CariesFG, and 3.77 % higher than OralNet. In terms of precision, NFR-EDL scores 91.60 %, marking improvements of 4.43 %, 2.22 %, and 4.40 %, respectively. Similarly, its average recall of 91.37 % surpasses that of ResNet34 (86.88 %) by 4.49 %, CariesFG (88.82 %) by 2.55 %, and OralNet (87.81 %) by 3.56 %. Furthermore, NFR-EDL achieves an average F1 Score of 91.4 %, outperforming CariesFG (89.00 %), ResNet34 (86.99 %), and OralNet (87.47 %) by 2.40 %, 4.41 %, and 3.93 %, respectively. While other models have shown strong performance, none were able to match the NFR-EDL model's overall accuracy and consistency. These findings emphasize the significant advancements made with the NFR-EDL model, which represents a notable improvement over existing techniques, offering higher accuracy, better balance between precision and recall, and greater overall diagnostic capability.

4.7. Statistical analysis

To validate the statistical significance of the results, a statistical evaluation is conducted comparing NFR-EDL with existing RGB photographing methods across 20 trials to ensure robust statistical evidence. Paired t-tests are employed to assess differences in accuracy performance between NFR-EDL and the baseline methods, given that accuracy provides a reliable estimate of overall model performance. To address the issue of multiple comparisons, the Bonferroni correction is applied, adjusting the significance threshold from $\alpha = 0.05$ to $\alpha = 0.0167$ (i.e., 0.05/3, based on three comparisons).

As presented in Table 6, all statistical tests reveal a significant improvement of NFR-EDL over the baseline techniques, even under the stricter threshold. The comparison with ResNet34 yields a highly significant result ($t = 8.43$, $p = 0.0023$), suggesting that the enhanced performance of NFR-EDL is unlikely to be due to random chance. In comparison to CariesFG, NFR-EDL demonstrates a notable improvement in accuracy ($t = 4.95$, $p = 0.0089$), despite CariesFG's competitive performance on specific data. The evaluation against OralNet also produces statistically significant outcomes ($t = 7.28$, $p = 0.0031$), further affirming the effectiveness of NFR-EDL. All comparisons result in p-values below the Bonferroni-adjusted significance level of 0.0167, confirming that the accuracy improvements shown in Table 5 are statistically meaningful, not due to random variation, and represent genuine advancements in classification accuracy. These findings strongly support the repeatability and reliability of NFR-EDL's superior

Table 6
Statistical significance analysis of NFR-EDL performance compared to baseline methods and existing techniques across all datasets (with Bonferroni correction, adjusted $\alpha = 0.0167$).

Model	p-value	t-statistic	Mean Difference	95 % Confidence Interval	Significant at adjusted $\alpha = 0.0167$
NFR-EDL vs ResNet34	0.0023	8.43	4.57 %	[2.83 %, 6.31 %]	Yes
NFR-EDL vs CariesFG	0.01	4.95	2.42 %	[1.18 %, 3.66 %]	Yes
NFR-EDL vs OralNet	0.0031	7.28	3.77 %	[2.21 %, 5.33 %]	Yes

performance across all datasets.

Further quantitative validation is provided through an analysis of performance distribution across iterations. Table 6 summarizes the statistical comparison between NFR-EDL and the baseline methods, incorporating the Bonferroni-corrected significance level ($\alpha = 0.0167$). The results show that the proposed NFR-EDL model achieves not only higher mean accuracy but also lower variability (standard deviation: 1.24 %) compared to the baseline models (standard deviations: ResNet34: 2.31 %, CariesFG: 1.87 %, OralNet: 2.05 %). This consistency indicates more stable performance across diverse data splits. Additionally, effect sizes are computed using Cohen's d, a standardized measure of mean differences. Substantial effect sizes are observed for all comparisons ($d = 1.97$ for NFR-EDL vs. ResNet34, $d = 1.29$ vs. CariesFG, and $d = 1.83$ vs. OralNet), underscoring the practical significance of the improvements beyond statistical validation.

5. Conclusion

In this paper, a Non-linear Fuzzy Rank-based Ensemble Deep Learning (NFR-EDL) model has been designed for the efficient diagnosis of oral and dental diseases. The proposed model integrates the strengths of multiple deep learning architectures, ensemble learning, and fuzzy techniques to enhance diagnostic accuracy and reliability. Through the generation of confidence scores by the fundamental models and subsequent mapping onto non-linear fuzzy rankings, an all-encompassing final score was meticulously computed to minimize any potential discrepancies and heighten the level of confidence in disease identification. Through extensive experiments on four datasets, the proposed NFR-EDL model consistently outperformed its individual base learners across various metrics including accuracy, precision, recall, and F1 score. In particular, the NFR-EDL model demonstrated superior performance in correctly classifying oral diseases, achieving high accuracy and robustness. The inclusion of a strict 94 % bootstrap confidence interval further validated the model's effectiveness, as reflected in the accuracy and loss curves, confusion matrices, and ROC curves for each dataset. This confirms that our model is well-trained and capable of delivering reliable results. Given the high prevalence and significant impact of periodontitis and dental caries on global health, our model offers a valuable tool for early detection and diagnosis, which is crucial for effective treatment and prevention.

Despite the advantages of the proposed NFR-EDL model, several limitations need to be addressed to further improve its applicability and performance. The datasets utilized in this study are from curated sources, which might not fully represent the complexity of real-world clinical data. Integrating data from actual dental clinics, which include a variety of noise and variability, could enhance the model's robustness and reliability in practical settings. Moreover, the current model focuses solely on image data. As a future work, integrating other types of data, such as textual information from patient records, clinical notes, and demographic data, could enhance diagnostic accuracy and provide a more holistic view of patient health. Another future work could explore the integration of additional deep learning techniques and the application of the model to a broader range of dental conditions to further enhance its diagnostic capabilities. Furthermore, we aim to explore the potential of contemporary architectures, such as transformers, to provide deeper insights into the trade-offs between model complexity, performance, and clinical applicability in identifying oral and dental diseases. By exploring these future directions, the NFR-EDL model can be refined and adapted to better serve dental professionals and improve patient outcomes in the diagnosis of oral and dental diseases.

CRediT authorship contribution statement

Pouyan Razmjouei: Writing – original draft, Visualization, Supervision, Methodology, Investigation, Data curation, Conceptualization. **Elaheh Moharamkhani:** Writing – original draft, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Seyed Sasan Aryanezhad:** Writing – review & editing, Supervision, Conceptualization. **Mohammad Shokouhifar:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Conceptualization. **Mehdi Hosseinzadeh:** Writing – original draft, Supervision, Project administration, Conceptualization. **Behrouz Zadmehr:** Writing – original draft, Visualization, Software, Investigation, Data curation.

Ethics statement

This study did not involve any direct experimentation on human subjects by the authors. Instead, the analysis was conducted using publicly available datasets, including Kaggle, MOD, ODSI-DB, and OaDD. These datasets are openly accessible for research purposes and have been anonymized by their respective providers to ensure the privacy and confidentiality of the individuals represented. As such, no additional ethical approval was required. All procedures were performed in compliance with relevant laws and institutional guidelines. Informed consent was obtained by the original data collectors, and the privacy rights of human subjects were strictly observed throughout this study.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to improve the language and readability of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. Yilmaz, M. Tasyurek, M. Amuk, M. Celik, E.M. Canger, Developing deep learning methods for classification of teeth in dental panoramic radiography, Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology 138 (1) (2024) 118–127.
[2] A. Falcao, P. Bullón, A review of the influence of periodontal treatment in systemic diseases, Periodontol 79 (1) (2019) 117–128, 2000.
[3] R.G. Watt, A. Heilmann, S. Listl, M.A. Peres, London charter on oral health inequalities, J. Dent. Res. 95 (3) (2016) 245–247.
[4] Y. Li, Y. Xiang, H. Ren, C. Zhang, Z. Hu, W. Leng, L. Xia, Association between periodontitis and dental caries: a systematic review and meta-analysis, Clin. Oral Invest. 28 (6) (2024) 1–15.
[5] K.-H. Yu, A.L. Beam, I.S. Kohane, Artificial intelligence in healthcare, Nat. Biomed. Eng. 2 (10) (2018) 719–731.
[6] F. Jiang, et al., Artificial intelligence in healthcare: past, present and future, Stroke Vasc Neurol 2 (4) (2017).
[7] C. Shen, D. Nguyen, Z. Zhou, S.B. Jiang, B. Dong, X. Jia, An introduction to deep learning in medical physics: advantages, potential, and challenges, Phys. Med. Biol. 65 (5) (2020).
[8] M.H. Mahoor, M. Abdel-Mottaleb, Classification and numbering of teeth in dental bitewing images, Pattern Recognit 38 (4) (2005) 577–586.
[9] H. Chen, et al., A deep learning approach to automatic teeth detection and numbering based on object detection in dental periapical films, Sci. Rep. 9 (1) (2019) 3840.
[10] K. Yoon, H.M. Jeong, J.W. Kim, J.H. Park, J. Choi, AI-based dental caries and tooth number detection, Int. Dent. J. 74 (2024) S213–S214.
[11] K. Yoon, H.M. Jeong, J.W. Kim, J.H. Park, J. Choi, AI-based dental caries and tooth number detection in intraoral photos: model development and performance evaluation, J. Dent. 141 (2024) 104821.
[12] I.D.S. Chen, C.-M. Yang, M.-J. Chen, M.-C. Chen, R.-M. Weng, C.-H. Yeh, Deep learning-based recognition of periodontitis and dental caries in dental x-ray images, Bioengineering 10 (8) (2023) 911.
[13] American Dental Association, Council on Scientific Affairs, The use of dental radiographs: Update and recommendations, J. Am. Dent. Assoc. 137 (9) (2006) 1304–1312.
[14] M. Rohlin, B. Kullendorff, M. Ahlqwist, C.O. Henrikson, L. Hollender, B. Stenström, Comparison between panoramic and periapical radiography in the diagnosis of periapical bone lesions, Dentomaxillofacial Radiol. 18 (4) (1989) 151–155.

[15] C.W. Douglass, R.W. Valachovic, A. Wijesinha, H.H. Chauncey, K.K. Kapur, B. J. McNeil, Clinical efficacy of dental radiography in the detection of dental caries and periodontal diseases, Oral Surgery, Oral Med Oral Pathol 62 (3) (1986) 330–339.
[16] R. Bouali, O. Mahboub, M. Lazaar, Review of dental diagnosis by deep learning models: Trends, applications and challenges, Procedia Comput. Sci. 231 (2024) 221–228.
[17] A. Broll, M. Goldhacker, S. Hahnel, M. Rosentritt, Generative deep learning approaches for the design of dental restorations: a narrative review, J. Dent. (2024) 104988.
[18] S.A. Prajapati, R. Nagaraj, S. Mitra, Classification of dental diseases using CNN and transfer learning, in: 2017 5th International Symposium on Computational and Business Intelligence (ISCBI), IEEE, 2017, pp. 70–74.
[19] K. Zhang, J. Wu, H. Chen, P. Lyu, An effective teeth recognition method using label tree with cascade network structure, Comput Med Imaging Graph 68 (2018) 61–70.
[20] C.-W. Li, et al., Detection of dental apical lesions using CNNs on periapical radiograph, Sensors 21 (21) (2021) 7049.
[21] M.G. Endres, et al., Development of a deep learning algorithm for periapical disease detection in dental radiographs, Diagnostics 10 (6) (2020) 430.
[22] Y.-C. Mao, et al., Deep learning for dental diagnosis: a novel approach to furcation involvement detection on periapical radiographs, Bioengineering 10 (7) (2023) 802.
[23] H. Shafiq, G. Gilanie, M. Sajid, M. Ahsan, Dental radiology: a convolutional neural network-based approach to detect dental disorders from dental images in a real-time environment, Multimed. Syst. 29 (6) (2023) 3179–3191.
[24] F. Liu, et al., Recognition of digital dental X-ray images using a convolutional neural network, J Digit Imaging 36 (1) (2023) 73–79.
[25] G.D. Deepak, S. Krishna Bhat, Optimization of deep neural networks for multiclassification of dental X-rays using transfer learning, Comput Methods Biomech Biomed Eng Imaging Vis 12 (1) (2024) 2272976.
[26] K. Kokomoto, R. Kariya, A. Muranaka, R. Okawa, K. Nakano, K. Nozaki, Automatic dental age calculation from panoramic radiographs using deep learning: a two-stage approach with object detection and image classification, BMC Oral Health 24 (1) (2024) 143.
[27] W. You, A. Hao, S. Li, Y. Wang, B. Xia, Deep learning-based dental plaque detection on primary teeth: a comparison with clinical assessments, BMC Oral Health 20 (2020) 1–7.
[28] D.M. Alalharith, et al., A deep learning-based approach for the detection of early signs of gingivitis in orthodontic patients using faster region-based convolutional neural networks, Int J Environ Res Public Health 17 (22) (2020) 8447.
[29] W. Li, Y. Liang, X. Zhang, C. Liu, L. He, L. Miao, W. Sun, A deep learning approach to automatic gingivitis screening based on classification and localization in RGB photos, Sci. Rep. 11 (1) (2021) 16831.
[30] A. Garg, J. Lu, A. Maji, Towards earlier detection of oral diseases on smartphones using oral and dental RGB images, arXiv Prepr arXiv230815705 (2023) 1–10. https://doi.org/10.48550/arXiv.2308.15705.
[31] S. Park, et al., Periodontal disease classification with color teeth images using convolutional neural networks, Electronics 12 (7) (2023) 1518.
[32] H. Jiang, P. Zhang, C. Che, B. Jin, Y. Zhu, CariesFG: a fine-grained RGB image classification framework with attention mechanism for dental caries, Eng. Appl. Artif. Intell. 123 (2023) 106306.
[33] C. Sulochana, M. Sumathi, Enhancing oral cancer diagnosis: IAWMF based preprocessing in RGB and CT images, in: 2024 International Conference on Recent Advances in Electrical, Electronics, Ubiquitous Communication, and Computational Intelligence (RAEEUCCI), IEEE, 2024, pp. 1–6.
[34] P. Sampath, N. Sasikaladevi, S. Vimal, M. Kaliappan, OralNet: deep learning fusion for oral cancer identification from lips and tongue images using stochastic gradient based logistic regression, Network Modeling Analysis in Health Informatics and Bioinformatics 13 (1) (2024) 24.
[35] Tooth condition data set." [Online]. Available: https://www.kaggle.com/datasets/salmansajid05/oral-diseases.
[36] J. Rashid, B.S. Qaisar, M. Faheem, A. Akram, R. ul Amin, M. Hamid, Mouth and oral disease classification using InceptionResNetV2 method, Multimed Tools Appl 83 (11) (2024) 33903–33921.
[37] J. Hyttinen, P. Fält, H. Jäsberg, A. Kullaa, M. Hauta-Kasari, Oral and dental spectral image database—odsi-db, Appl Sci 10 (20) (2020) 7246.
[38] J. Deng, A large-scale hierarchical image database, Proc IEEE Comput Vis Pattern Recognition (2009) 2009.
[39] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, 2014 *arXiv Prepr arXiv14091556*,.
[40] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[41] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[42] F.N. Iandola, S. Han, M.W. Moskewicz, K. Ashraf, W.J. Dally, K. Keutzer, SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size, arXiv Prepr arXiv160207360 (2016) 1–13. https://doi.org/10.48550/arXiv.1602.07360.
[43] H. Li, W. Li, J. Chang, L. Zhou, J. Luo, Y. Guo, Dermoscopy lesion classification based on GANs and a fuzzy rank-based ensemble of CNN models, Phys. Med. Biol. 67 (18) (2022) 185005.
[44] A. Manna, R. Kundu, D. Kaplun, A. Sinitca, R. Sarkar, A fuzzy rank-based ensemble of CNN models for classification of cervical cytology, Sci. Rep. 11 (1) (2021) 14538.
[45] S.D. Deb, R.K. Jha, Breast UltraSound Image classification using fuzzy-rank-based ensemble network, Biomed. Signal Process Control 85 (2023) 104871.