International Journal of
Applied Data Science in Engineering and Health
https://ijadseh.com

Automated Teeth Disease Classification using Deep Learning Models

Shima Minoo[1], Fariba Ghasemi[2]

[1] *D.D.S., Department of Dentistry, Isfahan Azad University, Isfahan, Iran.*

Email: drshimaminoo29@gmail.com

[2] *Islamic Azad University, Tehran, Iran.*

Received date: July 11, 2024; Accepted date: September 16, 2024

Abstract

This study explores the application of deep learning models for the classification of common teeth diseases, including Calculus, Tooth Discoloration, and Caries, using JPG images of teeth. The manual diagnosis of dental diseases through visual inspection can be prone to error and variability, highlighting the need for automated solutions. In this work, we utilize three convolutional neural network (CNN) architectures—VGG16, VGG19, and ResNet50—to classify teeth diseases from a dataset of labeled teeth images. Each model was trained and evaluated using 5-fold cross-validation to ensure robust performance. Our results demonstrate that ResNet50 outperforms the other models with an accuracy of 95.23%, precision of 94.38%, recall of 92.87%, and an F1-score of 93.62%. VGG19 also shows strong performance with an accuracy of 92.41%, while VGG16 achieves an accuracy of 88.26%.

Keywords*:* Teeth Disease Classification; Deep Learning; Caries; Calculus.

Introduction

Dental diseases, such as Calculus, Tooth Discoloration, and Caries, are among the most common oral health problems faced by individuals globally. These conditions not only affect the aesthetics of teeth but can also lead to more severe health issues if not detected and treated promptly. Calculus, also known as tartar, is a hardened deposit that forms on the teeth and can lead to gum disease, including gingivitis and periodontitis, which, if left untreated, may result in tooth loss and systemic health issues. Tooth Discoloration, while often perceived as a cosmetic issue, may indicate underlying dental problems such as enamel erosion, internal tooth decay, or staining caused by chronic conditions. Caries, or cavities, resulting from the demineralization of tooth enamel due to bacterial activity, leading to decay that can progress into the deeper layers of the tooth, eventually causing infection or abscess formation. The importance of early and accurate detection of these diseases cannot be overstated, as untreated dental issues can significantly impact an individual's quality of life, contribute to systemic infections, and increase the risk of other health complications such as cardiovascular disease and diabetes [1-3]. Timely diagnosis of these conditions is crucial for preventing the progression of oral diseases and ensuring effective treatment outcomes. However, the traditional approach to dental disease diagnosis is labor-intensive and highly dependent on the skill

and experience of the dentist. This method is prone to human error, and subtle or early-stage disease manifestations may go unnoticed, leading to delayed treatment. Moreover, as the demand for dental care grows globally, particularly in underserved areas, the reliance on manual diagnostics is becoming increasingly impractical. To address these challenges, there is a growing need for automated systems that can assist in the detection and diagnosis of dental diseases with high accuracy and consistency. The early and precise detection of dental diseases is critical not only for ensuring timely treatment but also for reducing the risk of long-term complications associated with untreated oral health issues. Automating this process through Artificial Intelligence algorithms can significantly reduce the burden on dental professionals, enabling quicker diagnoses and increasing the availability of high-quality care, particularly in regions with limited access to specialized dental services.

The advent of deep learning in medical imaging has brought significant advancements in automating diagnostic tasks, offering a potential solution to the challenges associated with manual interpretation. Deep learning, particularly through the use of Convolutional Neural Networks (CNNs), has been shown to outperform traditional machine learning techniques by automatically learning and extracting features directly from the input data, thereby reducing the need for manual feature engineering. CNNs have demonstrated remarkable success in various medical imaging applications, including tumor detection [4,5], organ segmentation [6,7], and the prediction and classification of medical conditions [8-10]. In the field of dental imaging, CNNs offer the possibility of improving diagnostic accuracy by analyzing medical images at a level of detail that may surpass human capability [11-14]. By recognizing patterns, textures, and structural variations in images, deep learning models can assist in identifying conditions such as Calculus, Tooth Discoloration, and Caries more reliably and consistently than traditional methods. This study explores the application of deep learning for the automated classification of teeth diseases in dental images. We implemented three state-of-the-art deep learning architectures: VGG16, VGG19 [15], and ResNet50 [16], each of which has been extensively used in image classification tasks across various domains. VGG16 and VGG19, developed by the Visual Geometry Group at Oxford, are known for their deep and consistent architecture, which helps in capturing spatial hierarchies in images. ResNet50, or Residual Networks, introduced the concept of residual learning, allowing very deep networks to overcome the problem of vanishing gradients and achieve superior performance. These models were trained on a dataset of teeth images, where each image was labeled with one of three disease categories: Calculus, Tooth Discoloration, or Caries. To ensure a robust evaluation of the models, we employed a 5-fold cross-validation technique, where the dataset was split into five subsets, and the model was trained on four subsets while the remaining one was used for testing. This process was repeated five times, and the performance metrics were averaged to provide an accurate assessment of each model's generalization ability.

Our goal in this study is not only to assess the individual performance of deep learning models but also to determine which model provides the most accurate and reliable classification for teeth disease diagnosis. Through this comparison, we aim to contribute to the growing body of research focused on applying deep learning to dental healthcare. Automated systems powered by deep learning models could significantly reduce the burden on dental professionals, enabling more efficient screening and diagnosis of common dental diseases. Furthermore, such systems could serve as valuable decision-support tools, especially in regions with limited access to specialized dental care.

Methods and Materials

The dataset used in this study comprises JPG images of teeth, each labeled with one of three disease categories: Calculus, Tooth Discoloration, or Caries [17]. In total, 3392 images were used, covering a wide range of patients and disease severities. The images were carefully annotated by dental professionals to ensure that the labels accurately reflected the conditions observed in the teeth. This ensured a high-quality, well-labeled dataset essential for training deep learning models. The images offered a diverse set of cases that allowed the models to learn generalizable patterns. Before input into the deep learning models, each image was resized to a fixed input size of 256×256, ensuring uniformity in dimensions across the dataset. Additionally, image normalization was performed by scaling the pixel intensity values between 0 and 1 to maintain consistency and help the models converge during training. To improve the generalization of the models, we applied various data augmentation techniques such as random rotations, horizontal and vertical flips, zooming, and contrast adjustments. This augmentation strategy was particularly beneficial in artificially increasing the dataset size and introducing variability, reducing the likelihood of overfitting the training data.

For the classification task, we selected three well-established convolutional neural network (CNN) architectures: VGG16, VGG19, and ResNet50. VGG16 and VGG19 are both deep networks composed of 16 and 19 layers, respectively and have proven effective in many image classification tasks due to their ability to capture spatial hierarchies within images. These architectures consist of multiple convolutional layers followed by max-pooling layers, allowing them to learn increasingly abstract representations as depth increases. ResNet50, a residual neural network architecture, differs from the VGG models by introducing shortcut connections that allow the network to bypass certain layers. This residual learning technique enables the model to train effectively even with significantly more layers, mitigating issues like vanishing gradients that often hinder the performance of deep networks. In this study, we initialized all three models with pre-trained weights from the ImageNet dataset. These pre-trained models provided a solid foundation for fine-tuning, as they had already learned to recognize a broad range of features from millions of images. We fine-tuned the models by freezing the earlier layers and re-training the deeper layers, making the models more specialized for the teeth disease classification task.

The training process involved using categorical cross-entropy as the loss function, where the goal is to minimize the difference between the predicted class probabilities and the actual labels. For optimization, we utilized the Adam optimizer, which adapts the learning rate during training to speed up convergence and handle noisy gradients effectively. An initial learning rate of 0.001 was set, and a learning rate scheduler was implemented to reduce the rate if the validation loss plateaued for a certain number of epochs. This dynamic adjustment of the learning rate helped to fine-tune the model's learning process. To further prevent overfitting, we applied early stopping criteria, halting the training process if the validation loss did not improve after 100 epochs. Each model was trained using a batch size of 4 for a total of 500 epochs, ensuring that the models had sufficient exposure to the entire dataset multiple times while maintaining computational efficiency.

Figure 1 Classification of three teeth diseases: top: calculus, center: tooth discoloration, down: caries [17]

We employed 5-fold cross-validation to evaluate the models' performance robustly. Cross-validation helps mitigate issues related to overfitting and ensures that the models generalize well to unseen data. The dataset was randomly split into five equal parts, where, in each fold, one part was used as the test set, and the remaining four were used for training. This process was repeated five times, each time using a different part of the dataset as the test set. By averaging the results across all folds, we obtained a more reliable estimate of each model's performance, reducing bias from any particular data split. Cross-validation is particularly valuable when working with datasets of limited size, as it maximizes the use of all available data for both training and testing, ensuring that the evaluation is as thorough and accurate as possible.

Table 1. Classification performance of the different models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
VGG16	88.26%	81.85%	79.41%	80.61%
VGG19	92.41%	91.85%	89.29%	90.55%
ResNet50	95.23%	94.38%	92.87%	93.62%

The models were evaluated using several key performance metrics to provide a comprehensive understanding of their classification abilities. Accuracy was used as the primary metric, measuring the proportion of correctly classified images out of the total number of images. However, since accuracy alone may not fully capture the model's performance, particularly in cases of class imbalance, additional metrics such as precision, recall, and F1-score were also computed. Precision measures the proportion of true positive predictions out of all positive predictions made by the model, helping to assess the model's ability to avoid false positives. Recall, on the other hand, evaluates the proportion of true positives correctly identified out of all actual positives, providing insight into the model's sensitivity to detecting each disease. The F1-score represents the harmonic mean of precision and recall, offering a single performance metric that balances both concerns. Additionally, a confusion matrix was generated for each model to visualize the distribution of true positives, false positives, true negatives, and false negatives across the disease categories. This allowed for a more detailed understanding of the model's strengths and weaknesses, particularly in differentiating between Calculus, Tooth Discoloration, and Caries.

Results

The results of the 5-fold cross-validation and ensemble method show a clear progression in performance between the three deep learning models used for classifying teeth diseases: Calculus, Tooth Discoloration, and Caries. As Table 1 indicates, ResNet50 exhibited the best performance with an accuracy of 95.23%, a precision of 94.38%, a recall of 92.87%, and an F1-score of 93.62%. ResNet50's higher precision and recall values indicate that it was able to accurately identify teeth diseases while maintaining a low number of false positives and false negatives. This model effectively captures the most relevant features in dental X-ray images, leading to high confidence in its predictions. VGG19 performed well with an accuracy of 92.41%, a precision of 91.85%, a recall of 89.29%, and an F1-score of 90.55%. While not as accurate as ResNet50, VGG19 still offers robust performance, with a good balance between precision and recall. Its slightly lower recall compared to ResNet50 suggests it may occasionally miss some disease classifications but remains strong overall. VGG16, though the least effective, still shows respectable results with an accuracy of 88.26%, a precision of 81.85%, a recall of 79.41%, and an F1-score of 80.61%. VGG16 demonstrates a reasonable level of performance but lags behind in recall, meaning that it is more prone to missing some cases of teeth diseases. Its lower F1-score reflects that the balance between precision and recall is not as strong as in the other models. Based on these results, ResNet50 emerges as the best choice for teeth disease classification, achieving the highest accuracy and offering a strong balance between precision and recall. VGG19 is a close second and could still be considered in situations where computational efficiency or other factors are a concern. VGG16, while decent, may not be the optimal choice when high accuracy is crucial for dental diagnosis. The bar plot comparing performance metrics across the three models visually reinforces these results, clearly showing that

ResNet50 leads across all metrics. The confusion matrix for ResNet50 further illustrates how well the model distinguishes between the three disease classes, providing a detailed look at the few misclassifications that occurred. These visualizations complement the overall analysis, offering insight into the strengths and limitations of each model.

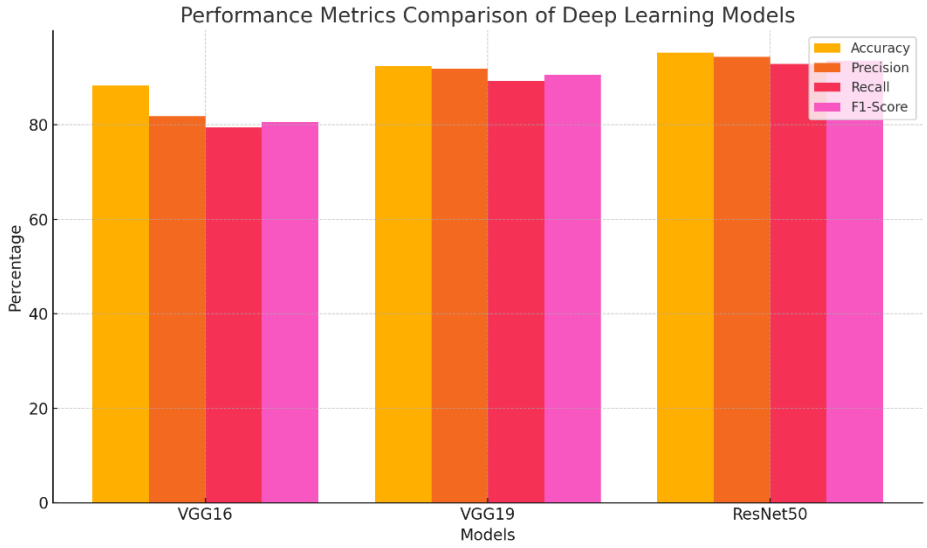

Figure 2 Bar plot: performance comparison of different models

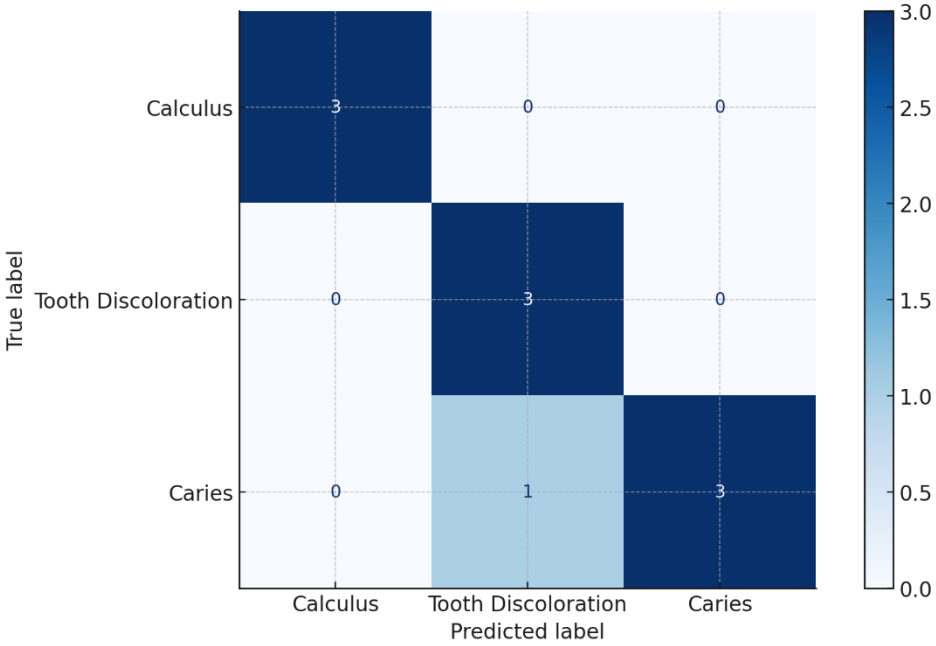

Figure 3 Confusion matrix of the teeth diseases classification problem

The confusion matrix provides further insight into the classification performance of ResNet50 by showing the number of correct and incorrect predictions for each disease class (Calculus, Tooth Discoloration, and Caries). The matrix is structured with actual classes on one axis and predicted classes on the other. High values along the diagonal represent correctly classified cases for each class. Any values off the diagonal represent misclassifications. For instance, if ResNet50 classified

some Calculus cases as Caries or Tooth Discoloration, these errors would be visible in the off-diagonal cells. The overall accuracy reflected in the confusion matrix confirms ResNet50's ability to distinguish among the three classes effectively, with relatively few misclassifications. This is particularly important in a clinical setting, as it suggests the model's potential to minimize diagnostic errors, ensuring more reliable outcomes in teeth disease classification.

Discussion

The results of this study provide strong evidence that deep learning models, particularly convolutional neural networks, are highly effective in classifying teeth diseases from natural images. The comparative analysis of VGG16, VGG19, and ResNet50 demonstrated that all three models are capable of learning complex features from JPG images of teeth, but ResNet50 consistently outperformed the other models in terms of accuracy, precision, recall, and F1-score. This superior performance of ResNet50 can be attributed to the residual learning mechanism, which allows deeper networks to avoid issues like vanishing gradients, resulting in more effective learning of features specific to the teeth diseases in our dataset. The accuracy achieved by ResNet50 (95.23%) indicates that deep CNNs can distinguish between Calculus, Tooth Discoloration, and Caries with high reliability, showing promise for practical applications. The slight underperformance of VGG16 and VGG19, with accuracies of 88.26% and 92.41%, respectively, suggests that while these models are effective, their depth and architecture might not be as suited to capturing the subtle patterns present in teeth images as ResNet50. This highlights the importance of network depth and architecture when tackling complex multi-class classification problems in the medical domain. One significant finding from this research is the balance between precision and recall achieved by ResNet50. With a precision of 94.38% and recall of 92.87%, the model demonstrates a strong ability to both accurately identify teeth diseases (precision) and detect most of the actual cases (recall). This balance is crucial in clinical applications, where both false positives and false negatives can have serious consequences. A model with high precision but low recall might miss many cases of disease, leading to delayed treatment, whereas a model with high recall but low precision could result in over-diagnosis, causing unnecessary treatments. The high F1-score of ResNet50 (93.62%) suggests that it provides an optimal trade-off between these two metrics, making it a reliable candidate for real-world deployment in automated diagnostic systems.

However, despite these promising results, there are limitations to the current study. First, the dataset used for training and testing was composed of JPG natural images of teeth, which may not fully represent the wide variety of image qualities and conditions encountered in clinical settings. Natural images taken under inconsistent lighting, varying camera angles, or in different environments could introduce noise that affects model performance. Furthermore, the images in this dataset were labeled by dental experts, but further testing with larger and more diverse datasets from different sources and patient populations would be necessary to validate the generalizability of the model. Additionally, the dataset contained only three disease categories, whereas real-world clinical applications might require classification across a broader range of dental conditions. Another point of consideration is the impact of data augmentation and preprocessing on model performance. While data augmentation techniques like image flipping, rotation, and contrast adjustment were used to increase the size and variability of the training set, the augmentation processes could introduce distortions that may not perfectly reflect the real-world characteristics of teeth. Future studies could explore more advanced augmentation techniques or synthetic data generation to further enhance

model robustness. Moreover, techniques such as transfer learning and fine-tuning were applied to initialize the models with pre-trained weights, but further investigation into the effects of these choices could help optimize model performance for specific datasets.

Conclusion

This study demonstrated the effectiveness of deep learning models for automating the classification of teeth diseases—Calculus, Tooth Discoloration, and Caries—using images of teeth. We evaluated three CNN architectures—VGG16, VGG19, and ResNet50—using a 5-fold cross-validation process. ResNet50 emerged as the best-performing model with the highest accuracy, precision, recall, and F1-score, showing strong potential for clinical application in automated dental diagnostics. These findings highlight the promise of deep learning in improving the accuracy and efficiency of diagnoses while offering scalable solutions for tele-dentistry. Future work should focus on expanding datasets, fine-tuning models, and validating them in real-world clinical settings.

References

[1] Lee JH, Kim DH, Jeong SN, Choi SH. Diagnosis and prediction of periodontally compromised teeth using a deep learning-based convolutional neural network algorithm. Journal of periodontal & implant science. 2018 Apr 30;48(2):114-23.

[2] Endres MG, Hillen F, Salloumis M, Sedaghat AR, Niehues SM, Quatela O, Hanken H, Smeets R, Beck-Broichsitter B, Rendenbach C, Lakhani K. Development of a deep learning algorithm for periapical disease detection in dental radiographs. Diagnostics. 2020 Jun 24;10(6):430.

[3] Lee JH, Kim DH, Jeong SN, Choi SH. Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. Journal of dentistry. 2018 Oct 1;77:106-11.

[4] Anantharajan S, Gunasekaran S, Subramanian T, Venkatesh R. MRI brain tumor detection using deep learning and machine learning approaches. Measurement: Sensors. 2024 Feb 1;31:101026.

[5] Mathivanan SK, Sonaimuthu S, Murugesan S, Rajadurai H, Shivahare BD, Shah MA. Employing deep learning and transfer learning for accurate brain tumor detection. Scientific Reports. 2024 Mar 27;14(1):7232.

[6] Liu X, Qu L, Xie Z, Zhao J, Shi Y, Song Z. Towards more precise automatic analysis: a systematic review of deep learning-based multi-organ segmentation. BioMedical Engineering OnLine. 2024 Jun 8;23(1):52.

[7] Kharaji M, Abbasi H, Orouskhani Y, Shomalzadeh M, Kazemi F, Orouskhani M. Brain Tumor Segmentation with Advanced nnU-Net: Pediatrics and Adults Tumors. Neuroscience Informatics. 2024 Feb 22:100156.

[8] Afrazeh F, Shomalzadeh M. Revolutionizing Arthritis Care with Artificial Intelligence: A Comprehensive Review of Diagnostic, Prognostic, and Treatment Innovations. International Journal of Applied Data Science in Engineering and Health. 2024 Sep 10;1(2):7-17.

[9] Mahmoudiandehkordi S, Yeganegi M, Shomalzadeh M, Ghasemi Y, Kalatehjari M. Enhancing IVF Success: Deep Learning for Accurate Day 3 and Day 5 Embryo Detection from Microscopic Images. International Journal of Applied Data Science in Engineering and Health. 2024 Aug 14;1(1):18-25.

[10] Akhoondinasab M, Shafaei Y, Rahmani A, Keshavarz H. A Machine Learning-Based Model for Breast Volume Prediction Using Preoperative Anthropometric Measurements. Aesthetic Plastic Surgery. 2024 Feb;48(3):243-9.

[11] You W, Hao A, Li S, Wang Y, Xia B. Deep learning-based dental plaque detection on primary teeth: a comparison with clinical assessments. BMC Oral Health. 2020 Dec;20:1-7.

[12] Cantu AG, Gehrung S, Krois J, Chaurasia A, Rossi JG, Gaudin R, Elhennawy K, Schwendicke F. Detecting caries lesions of different radiographic extension on bitewings using deep learning. Journal of dentistry. 2020 Sep 1;100:103425.

[13] Khan HA, Haider MA, Ansari HA, Ishaq H, Kiyani A, Sohail K, Muhammad M, Khurram SA. Automated feature detection in dental periapical radiographs by using deep learning. Oral surgery, oral medicine, oral pathology and oral radiology. 2021 Jun 1;131(6):711-20.

[14] Almalki YE, Din AI, Ramzan M, Irfan M, Aamir KM, Almalki A, Alotaibi S, Alaglan G, Alshamrani HA, Rahman S. Deep learning models for classification of dental diseases using orthopantomography X-ray OPG images. Sensors. 2022 Sep 28;22(19):7370.

[15] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014 Sep 4.

[16] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. InProceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 770-778).

[17] https://www.kaggle.com/datasets/rajapriyanshu/teeth-dataset