Journal of Electronics, Electromedical Engineering, and Medical Informatics
Multidisciplinary: Rapid Review: Open Access Journal **Vol. 7, No. 1, January 2025, pp: 117-129; eISSN: 2656-8632**

RESEARCH ARTICLE **OPEN ACCESS**

Manuscript received June 21, 2024; August 13, 2024; accepted August 20, 2024; date of publication November 20, 2024
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v7i1.487

How to cite: Yufis Azhar, Fauzan Adrivano Setiono, and Didih Rizki Chandranegara, "Comparison of Transfer Learning Models in Classification Dental and Tongue Disease Images", Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 7, no. 1, pp. 117-129, January 2025.

Comparison of Transfer Learning Models in Classification Dental and Tongue Disease Images

Yufis Azhar, Fauzan Adrivano Setiono, and Didih Rizki Chandranegara

Informatics Study Program, Faculty of Engineering, Universitas Muhammadiyah Malang, Indonesia

Corresponding author: Fauzan Adrivano Setiono (fauzanadrivano@webmail.umm.ac.id).

ABSTRACT According to the Global Burden of Disease Study, dental caries is the most prevalent oral health ailment, affecting around 3.5 billion individuals globally. According to the Ministry of Health of the Republic of Indonesia, 93% of children in the country suffer from oral health issues, making poor oral health a serious public health concern. The tongue and teeth in the mouth are particularly vulnerable to a wide range of illnesses, and the condition of the mouth is a key sign of the health of the body as a whole. The CNN algorithm has been utilized in numerous studies to classify disorders of the tongue and teeth. Nevertheless, no study has classified tongue and dental diseases using merged datasets as of yet. This research addresses this gap by focusing on the classification of dental and tongue diseases using transfer learning techniques with CNN architecture models VGG16, VGG19, and ResNet50. The primary aim is to compare these three models to identify the one with the most optimal performance in handling related cases. Based on the results, the best accuracy was achieved with data augmentation and models trained for 75 epochs. The VGG16 model attained 94% accuracy, VGG19 achieved 93% accuracy, and ResNet50 also reached 94% accuracy. These findings suggest that transfer learning with CNN architectures can effectively classify dental and tongue diseases. The implications are significant for developing automated diagnostic tools that can aid in the early detection and treatment of oral health issues globally.

INDEX TERMS CNN, VGG16, VGG19, ResNet50, Oral Diseases

I. INTRODUCTION

The oral cavity is the main gateway of the human system which consists of various biological niches such as teeth and tongue. Streptococcus Mutans is one of the many bacteria that predominate in the mouth cavity. These bacteria cause a clinical condition called dental caries [1]. Because oral bacteria make acid from carbohydrates, the surface of teeth becomes low pH, which leads to dental caries [2]. The disease causes white spot lesions that can develop into cavitation [3]. Apart from caries, there are various diseases that exist in the oral cavity, especially teeth and tongue. Dental diseases such as Caries, Calculus, Discoloration, and Gingivitis [4], tongue diseases such as Fissure, Geographic, and Black Hairy [5].

The Global Burden of Disease Study estimates that poor oral conditions affect nearly3.5 billion people in the world, with endless dental caries being the most common condition. An estimated 2.3 billion people suffer from endless carious teeth, and further than 530 million children suffer from primary carious teeth [6]. According to WHO, the prevalence of dental caries is 60-80% in children and nearly 100% in the adult population [7]. In Indonesia, poor oral health is a major public problem [8]. The Ministry Health of the Republic of Indonesia in 2018 stated that 93% of children experience oral problems [9]. The Basic Health Research shows that almost all provinces in Indonesia experienced a significant increase in the prevalence of active caries. The prevalence of active caries rose from 43.4% in 2007 to 53.2% in 2013 [10].

The high problem of oral disease encourages research to develop prediction systems using various Machine Learning algorithms. In the research of Saini et al. in 2021 [11]. With the title "Dental Caries early detection using Convolutional Neural Network for Tele dentistry", Convolutional neural

networks (CNN) like VGG16, VGG19, Inception V3, and Resnet50 are used to identify dental caries. Training, validation, and testing were performed on binary datasets with caries and non-caries images. The highest classification accuracy was achieved by the Inception V3 model, with a training accuracy of 99.89% and a validation accuracy of 98.95%. In another study by Alotaibi et al., in 2022 [12]. Which focused on developing an alveolar bone loss detection system on anterior periapical radiographs and categorizing the severity of bone loss due to periodontal disease. Deep CNN algorithms, especially VGG16, are useful in detecting elveolar bone loss and identifying the severity of bone damage. The accuracy results obtained were 73.0% in classifying normal vs disease, and 59% for classifying the severity of bone loss. Another study entitled "Classification of Approximal Caries in Bitewing Radiographs Using Convolutional Neural Networks" by Moran et al. in 2021 [13].

In order to detect approximate dental caries in bitewing radiography pictures and categorize them according to lesion severity, the study investigated the use of an artificial neural network and image processing system. During the 2000 iterations of the model training procedure, the Inception model produced the best results, with an accuracy of 73.0 and a learning rate parameter of 0.0001.

Another study related to dental classification was also conducted by Park et al., 2023 [14]. Entitled "Tooth caries classification with quantitative light-induced fluorescence (QLF) images using convolutional neural network for permanent teeth in vivo". Performing Quantitive Light-induced Fluorescence (QLF) image classification to detect dental caries using Convolutional Neural Network model. This study obtained results for the original QLF image with an accuracy of 83.2% while with QLF data that has been segmented it gets an accuracy of 85.6%. Shreyansh et al., 2018 [15], conducted research related to the classification of dental diseases with Radio Visiography (RVG) x-ray datasets. Classification is carried out using CNN architecture with the VGG16 model transfer learning technique and the best results are obtained in the CNN transfer learning scenario with an accuracy of 0.88% and CNN transfer learning with fine tuning with an accuracy of 0.88%.

This research is based on the understanding that Convolutional Neural Network (CNN) has a remarkable ability to extract hierarchical features from medical images based on previous literature studies. This ability is particularly relevant in diagnosing dental and tongue diseases, where subtle differences in texture, color, and shape can be important indicators of a disease. This research uses a transfer learning approach by comparing the performance of CNN models VGG16, VGG19, and ResNet50 in classifying dental and tongue health cases. In addition, the use of transfer learning is driven by its popularity and effectiveness in utilizing knowledge that has already been learned by existing models, thus reducing the time and resources required to train models from scratch. By utilizing the powerful feature extraction capabilities of CNNs, this study aims to develop a model that can accurately distinguish between healthy and infected tissues in the teeth and tongue. VGG architectures such as VGG16 and VGG19 were chosen in this study because VGG has a fairly simple and consistent architecture with the use of 3x3 convolution layers and 2x2 union layers. In addition, due to their popularity and effectiveness, VGG16 and VGG19 models are often used for research with transfer learning methods. ResNet50 was chosen for its residual architecture that introduces skip connections. This allows the model to overcome the vanishing gradient problem that often occurs in very deep networks. In addition, its 50 layers allow ResNet50 to capture more complex features from image data, as well as its good performance in dealing with large and complex datasets.

Previous studies have demonstrated the potential of convolutional neural networks (CNNs) in diagnosing oral diseases, especially dental diseases. However, to date, there has been no comprehensive study that directly compares the performance of the three models in classifying dental and tongue diseases. This prompts us to find out which CNN model is the most efficient in handling greater visual diversity on a combined tooth and tongue dataset.

This research aims to fill the gap by conducting an in-depth comparison between VGG16, VGG19, and ResNet50 in the context of tooth and tongue disease classification. Using a dataset that includes various types of diseases of the teeth and tongue, this study will identify which CNN model has the best performance in terms of accuracy, precision, and recall. We expect the results of this research to recommend the most suitable model for the development of a more accurate and efficient image-based oral disease diagnosis system. Knowledge of the most effective CNN models for diagnosing dental and tongue diseases can be utilized to develop better early diagnosis tools. This can help the Indonesian people to make early detection of oral diseases so that timely prevention and treatment can be taken, reducing the risk of more serious complications.

II. RESEARCH METHOD

FIGURE 1, displays the research flow conducted in this study. Starting from collecting the dataset needed in the research, then the data is divided into three subsets: train data, validation data, and test data. Next, the train and validation data are augmented to increase the amount and variety of data, besides that the augmentation process can also be used to reduce the possibility of overfitting in the training process. The classification models used are VGG16, VGG19, and ResNet50, and then the predetermined models are trained using training and validation data. The last is the evaluation stage which is carried out using test data that has never been seen by the model before, this process is useful for measuring and knowing the performance of the model objectively. The evaluation metrics used in this study include accuracy, precision, recall, and F1-score. These metrics provide a comprehensive insight into the performance of the model, thus enabling a thorough assessment of its effectiveness in classifying the target object.

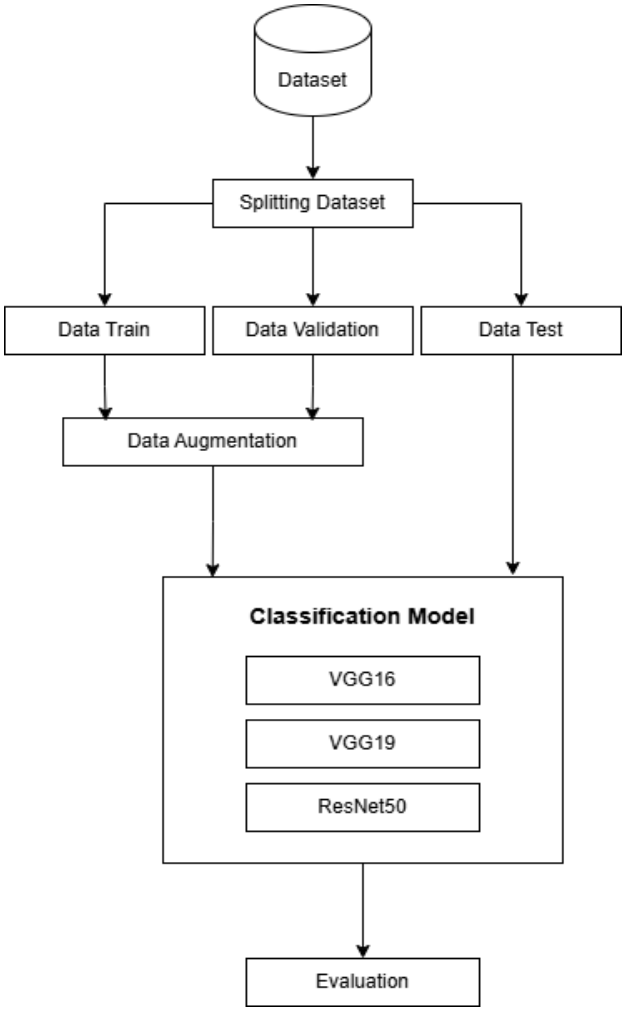


FIGURE 1. Research Flow

A. DATASET

For this study, a total of 8400 images were collected from several sources including Roboflow and Kaggle [16], [17], [18], [19], [20], [21], [22], [23], [24], [25]. This number of images was chosen after considering several factors, including the availability of data on these platforms and the need to maintain class balance in the dataset. Each class, for both teeth and tongue, consists of 700 images. This number of 700 was chosen as the minimum number representing the class with the least amount of data after all the data from various sources were combined and augmented. For other classes that have more data, an undersampling process is performed so that the distribution of data between classes is balanced. The distribution of all data is shown in TABLE 1.

TABLE 1
Data Proportion per Class

Class		Total
Teeth	Tongue	
Caries	Fissure	700
Calculus	Geographic	700
Discoloration	Diabetic	700
Gingivitis	Coated	700
Hypodontia	Black Hairy	700
Normal Teeth	Normal Tongue	700

Merging data from Roboflow and Kaggle aims to increase the diversity and complexity of the dataset. By combining data from various sources, it is expected that the model can learn more general and robust features. In addition, a careful data selection process was carried out to ensure the relevance of each image to its class. Duplicate or low-quality images were also removed to improve the quality of the dataset. This data collection process was carried out taking into account aspects such as image quality, pose variation, lighting, and patient condition to ensure a good representation of the actual patient population.

The data is divided into 3 subsets: with a proportion of 70% for training data, 20% for validation data, and 10% for test data. This resulted in a total of 5868 for train data, 1680 for validation data, and 852 for test data. Samples of the data used for each class can be seen in FIGURE 2.

B. DATA AUGMENTATION

Process augmentation is a technique in machine learning and data processing used to increase the amount and variety of training data without collecting new data. Variation in the dataset is necessary so that the model does not experience overfitting during the data training process. Data augmentation is usually applied to classification cases where there are class constraints learned from a defined table [26]. Data augmentation that is applied includes: rescale, shear range, zoom range, rotation range, horizontal & vertical flip,

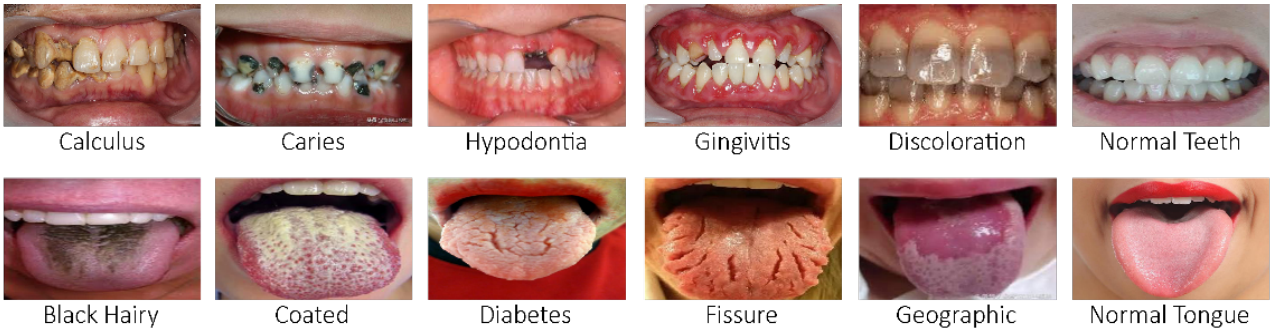


FIGURE 2. Samples Data for Every Class

and fill mode. The parameters of each augmentation process can be seen in TABLE 2.

TABLE 2
Data Augmentation

Augmentation	Settings
Rescale	1./255
Shear Range	0.2
Zoom Range	0.2
Rotation Range	45°
Horizontal & Vertical Flip	True
Fill Mode	Nearest

The augmentation process can also be used to balance the number of each class in the dataset. Some techniques in augmentation have proven useful in overcoming data imbalance [27]. Data imbalance in the dataset can lead to overfitting in the data training process. Details of the augmentation applied can be seen in table 2. The data augmentation process is applied to both training data and validation data. In addition to preventing overfitting, the data augmentation process can also help improve model performance in the learning process. The following is an example of the results of applying the augmentation process to the data shown in FIGURE 3.

FIGURE 3. Image Result After Augmentation Process

C. CONVOLUTIONAL NEURAL NETWORK

Convolutional Neural Network is a artificial neural network architecture used primarily in image processing and image analysis. CNN is currently the most effective model in classifying image data [28]. CNN is a deep learning network that can recognize and classify image features [29]. CNNs in particular have achieved successful results in medical image analysis and classification [30]. CNN has a convolution layer that allows the network to efficiently extract important features from images, such as edges, textures, and patterns. Convolutional neural networks (CNNs) have about fewer parameters and require very little data to process since they employ the original data as direct input to the network, eliminating the requirement for feature extraction or image reconstruction [31]. The main components of Convolutional Neural Network consist of convolutional layer, pooling layer, and fully connected layer, whose main function is to extract local features and detect normal distribution of computational information [32]. The architecture of CNN can be seen in FIGURE 4.

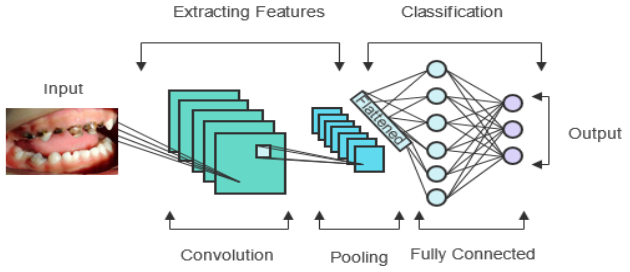


FIGURE 4. Convolutional Neural Network (CNN) Architecture

1) CONVOLUTIONAL

Convolutional layer is the core component of CNN, this layer plays an important role in extracting important features from data, especially data in images. The convolution operation equation between the input image and the filter can be seen in Eq. (1) [33].

$$0(x,y) = \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} I(x+i, y+i).F(i,j) \quad (1)$$

where $0(x,y)$ is an output feature map, $I(x+i, y+i)$ is input data/image, $F(i,j)$ is filters, and m, n is dimensions of filter

2) RELU ACTIVATION

ReLU is an activation function used in neural networks, it replaces negative values with zero and retains positive values. The ReLU activation equation can be seen in Eq. (2)

$$f(x) = \max(0, x) \quad (2)$$

where x is the variable function input, $f(x)$ is the output of the ReLU activation function for the input value x, where if the input is less than zero, the output will be zero. If the input is greater than or equal to zero, the output will be equal to itself.

3) POOLING (MAX POOLING)

Max pooling is a technique that exists in artificial neural networks, especially CNNs. This technique is useful for reducing the dimensionality of features and reducing the number of parameters that need to be learned. This process is done by taking the maximum value of a certain area in the feature map. The equation can be seen in Eq. (3) [34].

$$P_{j,m} = \max(h_{j,(m-1)N+r}) \quad (3)$$

where m-th max-polled band is composed of j related filters P_m, $N \in \{1,....,R\}$ is a pooling shift [35] .

4) FULLY CONNECTED

A fully connected layer is a layer in an artificial neural network that connects each neuron in the previous layer with each neuron in the next layer. The following equation is like Eq. (4) [33].

$$y = f\ (x.w) \tag{4}$$

where y is output vector, *f is* activation function, *x* is input vector, *w* is weight of the matrix

5) OUTPUT (SOFTMAX ACTIVATION)

Softmax activation is an activation function used in the last layer of a neural network to generate a probability distribution of possible classes. This function is suitable for use in multi-class classification problems. The softmax equation can be seen in Eq. (5) [36].

$$Softmax(z) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{zj}} \tag{5}$$

where softmax function $z = [z_1, z_2, \ldots z_n]$, z_i (is the score of class *i, e* is the Euler's number [35].

D. VGG16

VGG16 is one of the Convolutional Neural Network (CNN) architectures developed by the Visual Graphics Group (VGG) of Oxford University [37]. The VGGNet architecture is a developmental architecture of AlexNet that focuses on the feature extraction process in the convolutional layer, so as to be able to obtain multiple image representations for classification. VGG16 uses a relatively simple approach, where the architecture consists of a series of convolutions with 3 × 3 filters, has two consecutive filters to provide a more accessible 5 × 5 size field, and uses three 3 × 3 filters with a 7 × 7 filter size [38]. FIGURE 5, shown the architecture of the VGG16. However, vgg16 has a significant drawback in its high computational cost due to the large number of parameters, which can lead to longer training times and increased memory requirements.

E. VGG19

There are 19 layers in all in the VGG19 architecture, comprising 3 fully linked layers and 16 convolutional layers. The convolutional layer effectively uses a 3x3 filter with step 1, while the fully connected layer uses a 2x2 filter with step 2 [39]. VGG19 is known for its ability to extract deeper features from images, and the extensive use of 3x3 convolutional layers consequently helps in generating more complex feature representations.

However, like its predecessor, VGG16, VGG19 tends to be computationally expensive due to its depth and the large number of parameters. Additionally, the architecture's depth can contribute to overfitting, especially when dealing with smaller datasets. Despite this, the architecture has served as a foundational benchmark and has inspired subsequent improvements in the field of computer vision. The architecture of VGG19 can be seen in FIGURE 6.

F. RESNET50

ResNet50 is one of the Convolutional Neural Network (CNN) models introduced by Microsoft Research in 2015 [40]. With 50 inner layers that have been trained on at least one million photos from the ImageNet database, ResNet-50 is a version of the ResNet architecture [41]. This architecture belongs to the ResNet (Residual Network) and is well-known for its effective use in training very deep networks without experiencing performance degradation issues. ResNet50 is a type of ResNet that has 50 layers, consisting of 48 convolution layers, 1 maxpool layer, and 1 average pool layer [42].

However, despite its strengths, ResNet50 is not without its limitations. The model's depth can still lead to computational inefficiency, especially when deployed on resource-constrained devices. Additionally, while residual

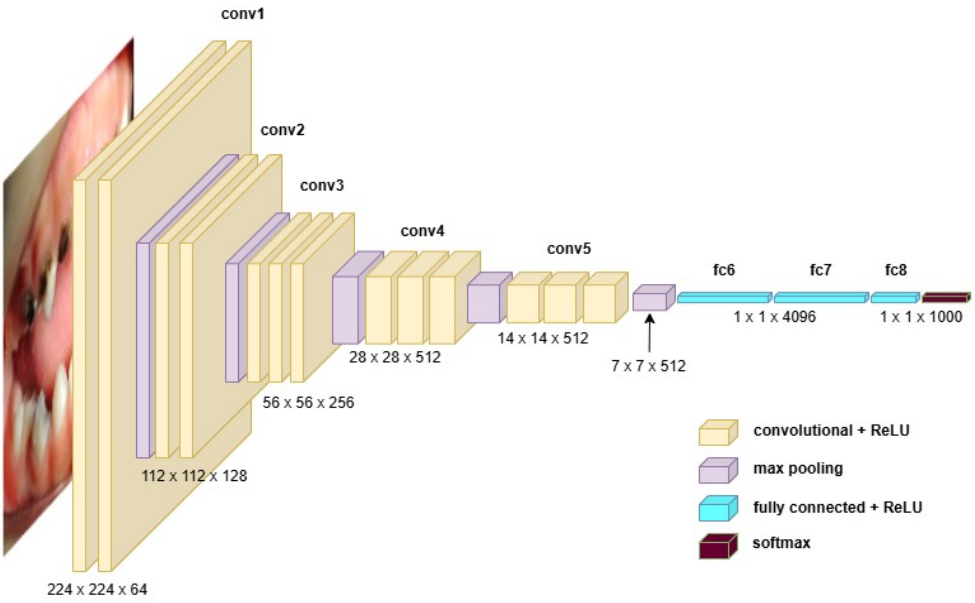


FIGURE 5. Architecture of VGG16

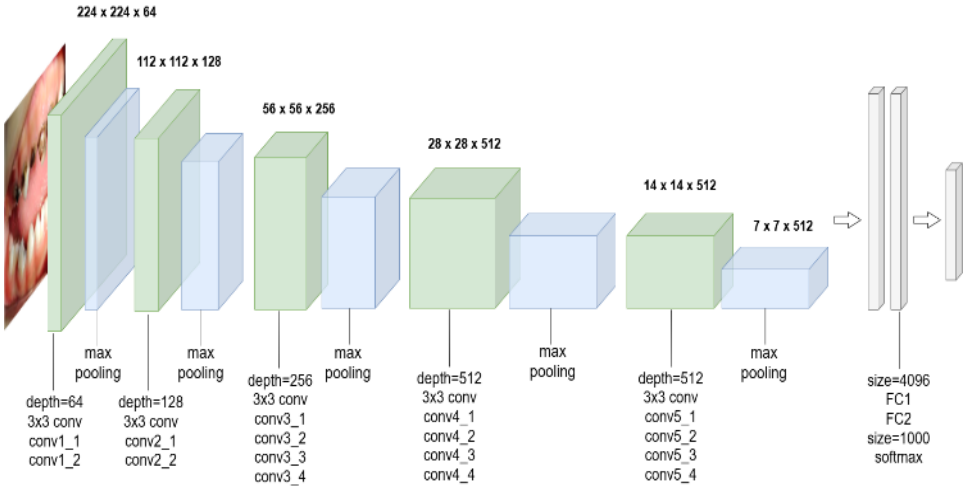


FIGURE 6. **Architecture of VGG19**

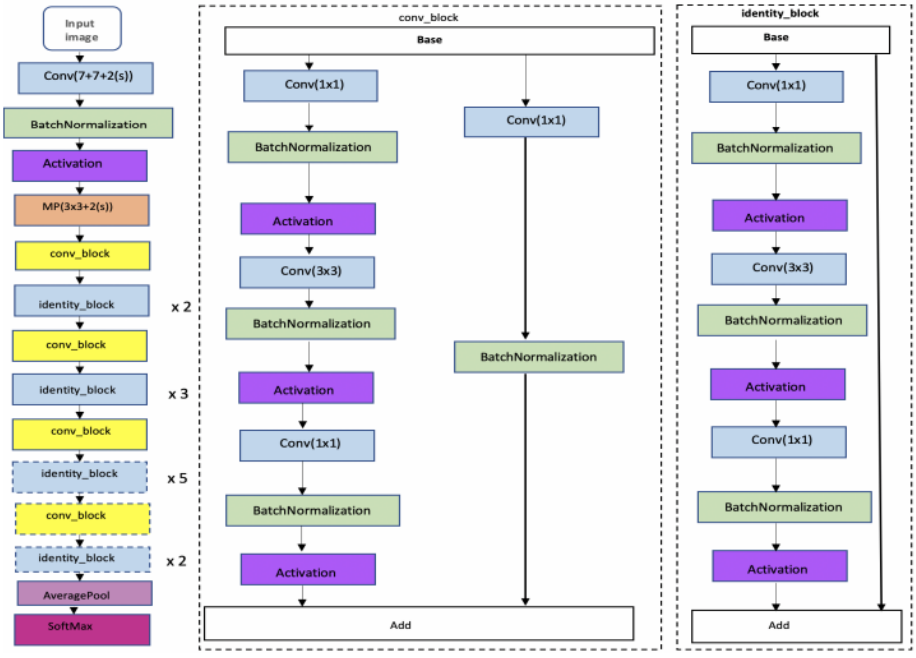


FIGURE 7. **Architecture of ResNet50**

connections help mitigate the vanishing gradient problem, they do not entirely eliminate it, and careful tuning of hyperparameters is often required to achieve optimal performance. FIGURE 7 shows the architecture of ResNet50 [43].

G. ARCHITECTURAL MODEL DESIGN

Parameter setting is crucial in machine learning because even a small adjustment can significantly affect the performance of the model. Many parameters can be used and customized according to the research needs. In this study, to identify the most optimal parameter values, the author conducted experiments by trying various combinations until the most optimal combination was found according to the author. The values used as selected parameters are shown in TABLE 3.

TABLE 3
Parameter User for Training Model

Parameter	Settings
Loss	Categorical Crossentrophy
Optimizer	SGD
Learning Rate	0.0001
Momentum	0.9
Step per Epoch	100
Verbose	1

This research uses the transfer learning method with the VGG16, VGG19, and ResNet50 model architectures. The three models use the same composition to classify dental and tongue diseases. As the input layer is adjusted to the model used with a size of 224x224, the basic model used is a model that has been trained with weights from the imagenet dataset, with tuning models such as removing the top layer (fully connected layers), adding convolution layer (Conv2D), pooling layer (MaxPooling2D), flatten layer, dense layer: 512-unit dense and 12-unit dense, and output layer (Softmax Activation). The results of the summary model with this composition can be seen in TABLE 4, TABLE 5, and TABLE 6.

TABLE 4
Model Summary of VGG16

Type of Layer	Output Shape	Param#
vgg16 (Functional)	(None, 7, 7, 512)	14,714,688
conv2d (Conv2D)	(None, 7, 7, 64)	294,976
max_pooling2d (MaxPooling2D)	(None, 3, 3, 64)	0
flatten (Flatten)	(None, 576)	0
dense (Dense)	(None, 512)	295,424
dense_1 (Dense)	(None, 12)	6,156

TABLE 5
Model Summary of VGG19

Type of Layer	Output Shape	Param#
vgg19 (Functional)	(None, 7, 7, 512)	20,024,384
conv2d (Conv2D)	(None, 7, 7, 64)	294,976
max_pooling2d (MaxPooling2D)	(None, 3, 3, 64)	0
flatten (Flatten)	(None, 576)	0
dense (Dense)	(None, 512)	295,424
dense_1 (Dense)	(None, 12)	6,156

TABLE 6
Model Summary of ResNet50

Type of Layer	Output Shape	Param#
resnet50 (Functional)	(None, 7, 7, 2048)	23,587,712
conv2d (Conv2D)	(None, 7, 7, 64)	1,179,712
max_pooling2d (MaxPooling2D)	(None, 3, 3, 64)	0
flatten (Flatten)	(None, 576)	0
dense (Dense)	(None, 512)	295,424
dense_1 (Dense)	(None, 12)	6,156

H. EVALUATION

In this research, we compared the performance of three deep learning models, namely VGG16, VGG19, and ResNet50, for classifying dental and tongue diseases. To evaluate the performance of each model, we employed a confusion matrix. The confusion matrix allows us to calculate various metrics such as accuracy, precision, recall, and F1-score.

These metrics are crucial, as they provide a comprehensive overview of the model's performance. Accuracy indicates the overall proportion of correct predictions. Precision measures the proportion of positive predictions that are actually correct, while recall measures the proportion of actual positive cases that are correctly identified. The F1-score is the harmonic mean of precision and recall, providing a balance between the two.

By analyzing these metrics, we can identify which model is most effective in classifying dental and tongue diseases. For example, if a model has high accuracy but low precision, it frequently classifies negative instances as positive. Conversely, if recall is low, the model often fails to identify positive instances. Therefore, selecting the appropriate metrics is crucial in evaluating classification models. The following is the formula for calculating the confusion matrix used:

1) ACCURACY

Accuracy is the ratio of correctly classified data to total observations [44]. The following equation is like Eq. (6):

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

2) PRECISION

Precision is the ratio of correctly predicted positive results to total predicted positive results [45]. The following equation is like Eq. (7):

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

3) RECALL

Recall is the ratio of correct positive results to total observations in the class, indicating the proportion of positive observations [46]. Recall is used to find out how many of all samples that should be positive and successfully predicted as positive by the model. The following equation is like Eq. (8):

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

4) F1-SCORE

F1-Score is the harmonic mean value of precision and recall [47]. The following equation is like Eq. (9):

$$F1 - Score = 2 . \frac{Precision . Recall}{Precision+Recall} \quad (9)$$

TP (True Positive) is the correct prediction for the positive class, TN (True Negative) is the correct prediction for the negative class, FP (False Positive) is a false prediction for a positive class, and FN (False Negative) is a false prediction for the negative class.

III. RESULT

This study conducted tests using 3 models such as VGG16, VGG19, and ResNet50 for the case of tooth and tongue disease classification. The result of this research is the categorization of dental and tongue diseases using the VGG16, VGG19, and ResNet50 models. Tests were conducted on the three models to compare the best performance results related to the classification of dental and tongue diseases. Where the assessment parameters include

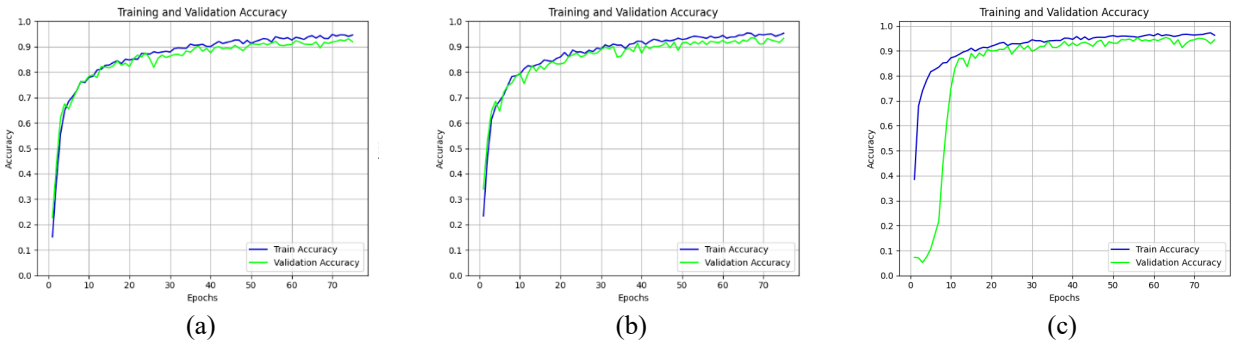


FIGURE 8. Training Graphic of a) VGG16, b) VGG19, and, c) ResNet50

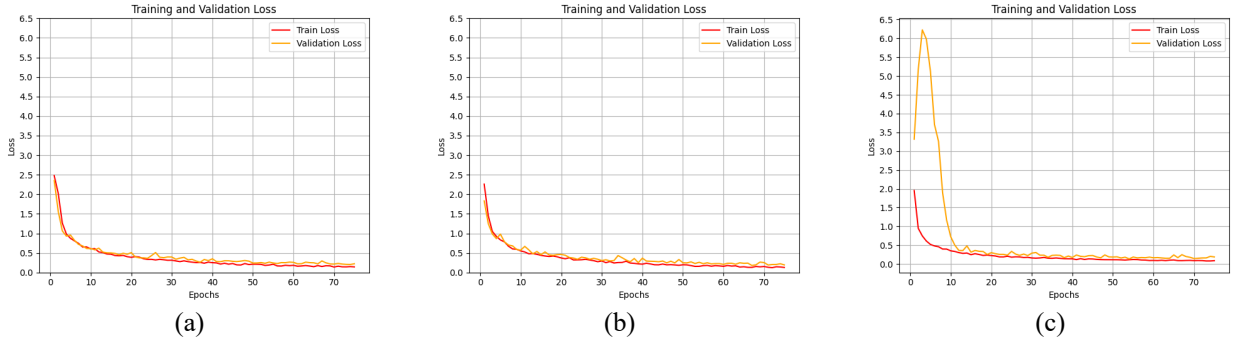


FIGURE 9. Loss Graphic of a) VGG16, b) VGG19, and, c) ResNet50

precission, recall, f1-score, and accuracy. In addition, this study also compares the prediction results of the categories performed by a model using confusion matrix. Testing of the three models was carried out using the same epoch of 75.

A. COMPARISON OF TRAIN AND LOSS ACCURACY

Based on FIGURE 8, the results after training with 75 epochs, all models show good performance. In general, the results achieved are quite satisfactory, indicating that the training process is further able to improve the predictive ability of the three models.

However, there is an important note regarding the ResNet50 model which experienced the most significant overfitting compared to the other models. Despite achieving good results at the end of the training process, this improvement should be weighed against the risk of overfitting which could reduce the model's ability to generalize new patterns.

ResNet50 tends to experience overfitting at the beginning of the training process because the deep structure with many layers may cause the model to learn small details of complex, even irrelevant training data more easily, resulting in overfitting of the data. However, as the number of epoch increases, ResNet is able to adjust and capture more complex patterns, so its performance becomes better and overfitting is reduced. Meanwhile, VGG tends to be more stable at the beginning of the training process because it has a simpler structure than ResNet, making it less prone to overfitting. However, despite being stable, the final result of VGG was still slightly below or on par with ResNet, indicating that despite the initial overfitting, ResNet was able to catch up and even surpass the performance of VGG at the end of the training process due to its ability to capture more complex patterns.

From the analysis of the loss graphs for the three models in FIGURE 9, it can be seen that VGG shows a stable loss rate at the beginning of the training process, showing stability in fitting the patterns from the training data. On the other hand, ResNet experiences significant overfitting at first, with a very high loss rate at the beginning which then drops rapidly. However, as the number of epochs increased, ResNet was able to adjust and lower its loss, demonstrating its ability to catch up and even surpass the performance of VGG in the later stages of training. In conclusion, although ResNet experienced initial challenges in the form of overfitting, its ability to improve over time makes it a strong choice, especially if the main focus is the final performance of the model.

B. COMPARISON OF CONFUSION MATRIX

From the confusion matrix results in FIGURE 10, it can be seen that the VGG16 model has the lowest prediction error compared to VGG19 and ResNet50. This shows that VGG16 is able to classify the data more accurately, with a lower error rate. However, VGG19 and ResNet50 showed slightly higher error rates compared to VGG16. Nonetheless, there was a consistent pattern across the three models, with the most incorrect guesses occurring in the Gingivitis and Calculus classes. The main cause of the high number of incorrect guesses in the Gingivitis and Calculus classes may be related to the complexity and variation in the characteristics of these two conditions.

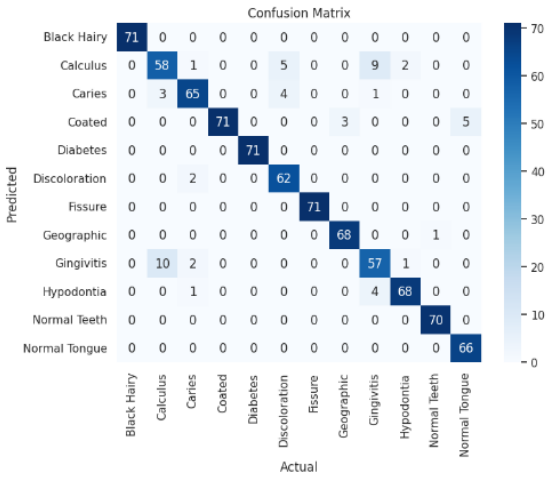
Gingivitis, which is an inflammation of the gums, and Calculus, which is a buildup of hardened plaque on the teeth, have characteristics that vary from case to case, making it difficult for the model to recognize consistent patterns within these classes. In addition, confusion between these two classes could also occur due to similar characteristics, especially in the context of the dataset used. Another possibility is that these two conditions may have visual representations that are more difficult to clearly distinguish in the image data, making it difficult for the model to distinguish between the Gingivitis and Calculus classes.

Based on the confusion matrix results, the VGG16 model shows better performance than VGG19 and ResNet50 in classifying the data. This can be seen from the lower prediction error rate. The superiority of VGG16 is likely due to its simpler architectural structure, yet it is effective in extracting relevant features for tooth and tongue condition classification. However, it also had difficulty distinguishing between the gingivitis and calculus classes, suggesting that even the best-performing models still face challenges in dealing with the visual complexity of these conditions.

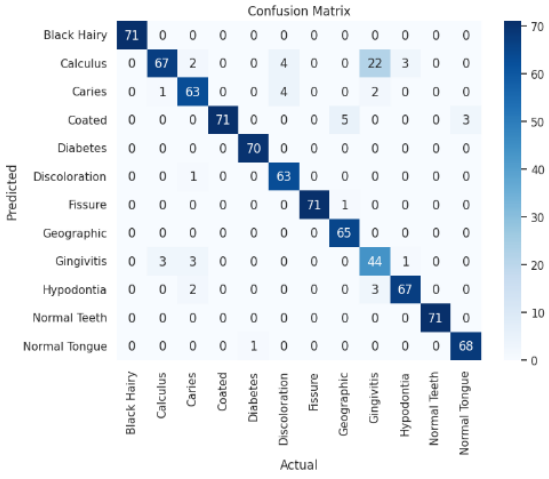
The VGG19 and ResNet50 models, despite having more complex architectures than VGG16, showed slightly higher error rates. This could be due to several factors. Firstly, the complexity of the architecture may increase the potential for overfitting, where the model overfits itself to the training data and thus is less able to generalize to new data. Secondly, these models may require a larger amount of data to achieve optimal performance. Finally, differences in architecture design, such as the use of residual connections in ResNet50, may affect the model's ability to extract certain features.

C. COMPARISON OF EVALUATION METRICS

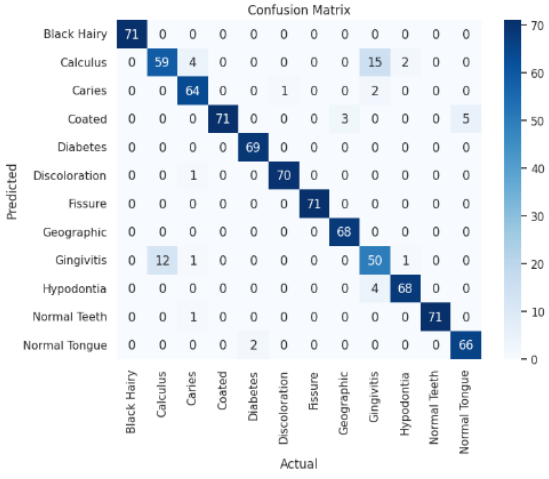
Based on the metric evaluation results of the three different models in TABLE 7, TABLE 8, and TABLE 9, VGG16, VGG19, and ResNet50, It can be seen that VGG16 and ResNet50 generally perform very well in most classes, with high F1-Score, especially in classes such as Caries, Discoloration, Hypodontia, and Normal Teeth. ResNet50 had an edge in overall performance with the highest F1-Score in some classes such as Discoloration and Gingivitis, while VGG16 was also consistently high in many classes. VGG19 showed good results, but was slightly inferior to the other two models, especially in the Gingivitis and Calculus classes.



(a)



(b)



(c)

FIGURE 10. Confusion Matrix of a) VGG16, b) VGG19, and, c) ResNet50

The results of this study demonstrate the potential of using deep learning models, particularly VGG16, as diagnostic aids in dentistry. The model's ability to classify tooth and tongue diseases with high accuracy can contribute to the early detection of diseases, thus enabling more effective medical interventions. However, the challenges in classifying gingivitis and calculus diseases indicate the need for further research to improve the model's performance in detecting these conditions. In addition, exploration of more complex model architectures and the use of more advanced data augmentation techniques could be the focus of future research.

TABLE 7
Evaluation Metrics Model of VGG16

Class	Precission (%)	Recall (%)	F1-Score (%)	Accuracy (%)
Caries	92	89	90	
Calculus	82	77	79	
Discoloration	87	97	92	
Gingivitis	80	81	81	
Hypodontia	96	93	94	
Normal Teeth	99	100	99	94
Fissure	100	100	100	
Geographic	96	99	97	
Diabetic	100	100	100	
Black Hairy	100	100	100	
Normal Tongue	93	100	96	

TABLE 8
Evaluation Metrics Model of VGG19

Class	Precission (%)	Recall (%)	F1-Score (%)	Accuracy (%)
Caries	89	90	89	
Calculus	94	68	79	
Discoloration	89	98	93	
Gingivitis	62	86	72	
Hypodontia	94	93	94	
Normal Teeth	100	100	100	93
Fissure	100	99	99	
Geographic	92	100	96	
Diabetic	99	100	99	
Black Hairy	100	100	100	
Normal Tongue	96	99	97	

TABLE 9
Evaluation Metrics Model of ResNet50

Class	Precission (%)	Recall (%)	F1-Score (%)	Accuracy (%)
Caries	90	96	93	
Calculus	83	74	78	
Discoloration	99	99	99	
Gingivitis	70	78	74	
Hypodontia	96	94	95	
Normal Teeth	100	99	99	94
Fissure	100	100	100	
Geographic	96	100	98	
Diabetic	97	100	99	
Black Hairy	100	100	100	
Normal Tongue	93	97	95	

IV. DISCUSSION

This study aimed to evaluate the performance of VGG16, VGG19, and ResNet50 in classifying dental and oral conditions using deep learning. Our results demonstrate that these models exhibit comparable performance, achieving accuracy ranging from 93% to 94%. While there are slight variations in the F1-score, the overall high accuracy suggests that these models are suitable for this task. Therefore, model selection can be further refined based on factors such as model complexity and computational resources.

Based on the analysis of train loss graph, confusion matrix, and evaluation metrics, this study has provided a comprehensive overview of the performance of three different models, namely VGG16, VGG19, and ResNet50, in classifying dental and tongue images. VGG16 showed consistent stability in prediction with a low error rate, while VGG19 and ResNet50 showed a higher ability in capturing complex patterns, albeit with a slightly higher error rate. Nonetheless, the main challenge was seen in the three models' difficulty in classifying the Gingivitis and Calculus classes, indicating the need for further improvement in recognizing these conditions more accurately.

In addition, graphical analysis shows that ResNet50 tends to experience overfitting early in the training process, characterized by a rapid decrease in train loss. The deep structure of ResNet tends to cause overfitting in the early stages of training, where the large number of layers allows the model to capture small details of complex training data, even those that may be irrelevant, resulting in overfitting of the data.

Previous research has established the efficacy of CNN models, particularly VGG16, in detecting dental diseases using radiographic images. However, these studies have primarily focused on dental conditions, neglecting the potential of image analysis for diagnosing tongue diseases. Our research expands upon this foundation by incorporating tongue image analysis into the diagnostic process.

By comparing the performance of VGG16, VGG19, and ResNet50 architectures, our study offers a comprehensive evaluation of deep learning models for oral disease detection. Our findings demonstrate a significant improvement in accuracy, with VGG16 and ResNet50 achieving a peak performance of 94%. These results underscore the potential of our approach to enhance the early detection and diagnosis of oral health conditions.The following comparison between the best results of previous research and current research will be shown in TABLE 10.

TABLE 10
Result of Previous Research and Current Research

	Model	Accuracy
Previous Research [15]	Transfer Learning VGG16	88%
Current Research	Transfer Learning VGG16	94%

Despite its strong performance, ResNet50 showed signs of overfitting early in the training process, as indicated by the rapid decrease in train loss. This could limit the model's ability to generalize well to unseen data, suggesting that future research should consider regularization techniques or

adjustments to training strategies to mitigate overfitting. The computational resources used in this study may not be sufficient to handle larger datasets or more complex models efficiently. Training deep learning models like VGG16, VGG19, and ResNet50 on larger datasets requires high-performance hardware (e.g., GPUs or TPUs). This limitation restricts the scalability of the study and prevents further exploration of more complex architectures or larger, more diverse datasets.

V. CONCLUSION

This study successfully provides a comprehensive overview of the performance of three image classification models, namely VGG16, VGG19, and ResNet50, in the context of tooth and tongue image classification. The analysis shows that VGG16 and VGG19 have consistent stability during the training process, while VGG16 and ResNet50 get the best results in predicting with the lowest error value. Compared to the other two models, VGG16 is the most effective model for classifying cases of tooth and tongue diseases, according to the tests that have been conducted. VGG16 has good stability and final value with an accuracy rate of 94%, while VGG19 obtained an accuracy value of 93%, and ResNet50 with an accuracy of 94%. Although ResNet50 had the same accuracy as VGG16, it experienced instability or overfitting during the training process at the beginning of the iteration. The main challenge faced was the difficulty of the three models in accurately classifying the Gingivitis and Calculus classes. In addition, ResNet50 tended to experience overfitting early in the training process, indicating the need for a better regulation strategy to overcome this overfitting. Overall, this study indicates that while each model has its strengths and weaknesses, there is significant potential to improve model performance through further customization of training techniques and data augmentation.

The results of this study contribute to the growing body of knowledge on applying deep learning to medical image analysis. By demonstrating the potential of these models for oral disease classification, this research lays the groundwork for future studies to explore more complex architectures, larger datasets, and advanced techniques to enhance diagnostic accuracy and clinical utility.

For future research, several adjustments and improvements should be considered. Firstly, the addition of more varied and representative training data could help improve the model's performance in classifying difficult classes such as Gingivitis and Calculus. Secondly, this research can be extended by considering the use of more sophisticated data augmentation techniques to increase the diversity of the training data and prevent overfitting. In addition, exploration of more advanced model optimization techniques, as well as more careful parameter adjustments, can help improve the overall performance of the model. Finally, to strengthen the results, model evaluation can also be extended to include other metrics such as area under curve (AUC) and the use of cross-validation techniques to validate the stability and consistency of model performance more broadly.

REFERENCES

[1] V. Ranganathan and C. Akhila, "Streptococcus mutans: has it become prime perpetrator for oral manifestations?," *J. Microbiol. Exp.*, vol. 7, no. 4, pp. 206–213, 2019.

[2] M. Sotozono *et al.*, "Impact of sleep on the microbiome of oral biofilms," *PLoS One*, vol. 16, no. 12 December, 2021.

[3] M. A. R. Buzalaf *et al.*, "Saliva as a diagnostic tool for dental caries, periodontal disease and cancer: is there a need for more biomarkers?," *Expert Rev. Mol. Diagn.*, vol. 20, no. 5, pp. 543–555, 2020.

[4] A. M. Agbor and Y. F. J. Jupkwo, "Oral Health of Tobacco and Non-Tobacco Consumers Inyaounde, Cameroon," *Eur. J. Dent. Oral Heal.*, vol. 1, no. 2, 2020.

[5] F. della Vella *et al.*, "The pseudolesions of the oral mucosa: Differential diagnosis and related systemic conditions," *Appl. Sci.*, vol. 9, no. 12, pp. 1–8, 2019.

[6] D. C. Experience, A. Factors, and A. Bahar, "Journal of International Dental and Medical Research ISSN 1309-100X http://www.jidmr.com Dental Caries Experience and Associated Factors Armasastra Bahar and et al," pp. 666–670, 2021.

[7] W. Qiu *et al.*, "Application-of-AntibioticsAntimicrobial-Agents-on-Dental-CariesBioMed-Research-International.pdf," *Biomed Res. Int.*, vol. 20, no. 20, pp. 1–11, 2020.

[8] C. M. A. Santoso, T. Bramantoro, M. C. Nguyen, Z. Bagoly, and A. Nagy, "Factors affecting dental service utilisation in indonesia: A population-based multilevel analysis," *Int. J. Environ. Res. Public Health*, vol. 17, no. 15, pp. 1–11, 2020.

[9] A. Riolina, S. Hartini, and S. Suparyati, "Dental and oral health problems in elementary school children: A scoping review," *Pediatr. Dent. J.*, vol. 30, no. 2, pp. 106–114, 2020.

[10] I. Dewanto, S. Koontongkaew, and N. Widyanti, "Characteristics of Dental Services in Rural, Suburban, and Urban Areas Upon the Implementation of Indonesia National Health Insurance," *Front. Public Heal.*, vol. 8, no. May, pp. 1–8, 2020.

[11] D. Saini, R. Jain, and A. Thakur, "Dental Caries early detection using Convolutional Neural Network for Tele dentistry," *2021 7th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2021*, pp. 958–963, 2021.

[12] G. Alotaibi, M. Awawdeh, F. F. Farook, and M. Aljohani, "Artificial intelligence (AI) diagnostic tools : utilizing a convolutional neural network (CNN) to assess periodontal bone level radiographically — a retrospective study," pp. 1–7, 2022.

[13] M. Moran, M. Faria, G. Giraldi, L. Bastos, L. Oliveira, and A. Conci, "Classification of approximal caries in bitewing radiographs using convolutional neural networks," *Sensors*, vol. 21, no. 15, pp. 1–12, 2021.

[14] E. Y. Park, S. Jeong, S. Kang, J. Cho, J. Cho, and E. Kim, "Tooth caries classification with quantitative light-induced fluorescence (QLF) images using convolutional neural network for permanent teeth in vivo," pp. 1–8, 2023.

[15] S. A. Prajapati, R. Nagaraj, and S. Mitra, "Classification of Dental Diseases U sing CNN and Transfer Learning Shreyansh," *Proc. - 6th Int. Symp. Comput. Bus. Intell. ISCBI 2018*, 2018.

[16] S. Sajid, "Oral Diseases," *Kaggle*. [Online]. Available: https://www.kaggle.com/datasets/salmansajid05/oral-diseases.

[17] Tongue Classification V2, "fissured tongue Dataset," *Roboflow*, 2023. [Online]. Available: https://universe.roboflow.com/tongue-classification-v2/fissured-tongue.

[18] Makhluk, "Black Hairy Tongue Dataset," *Roboflow*, 2023. [Online]. Available: https://universe.roboflow.com/makhluk/black-hairy-tongue.

[19] T. C. V2, "2black Dataset," *Roboflow*, 2023. [Online]. Available: https://universe.roboflow.com/tongue-classification-v2/2black.

[20] J. Dabass, "Tongue-coating," *Kaggle*. [Online]. Available: https://www.kaggle.com/datasets/jyotidabas/tongue-coating/data.

[21] T. Towfiq and Geeksforkicks, "Tongue__DIABETES," *Kaggle*. [Online]. Available: https://www.kaggle.com/datasets/towfiqtomal/tongue-diabetes.

[22] Skripsi191402083, "Diabetes Identification by Tongue Images Dataset," *Roboflow*, 2023. [Online]. Available: https://universe.roboflow.com/skripsi191402083/diabetes-identification-by-tongue-images.

[23] G. Tongue, "geographic tongue Dataset," *Roboflow*, 2023. [Online].

Available: https://universe.roboflow.com/gepgraphic-tongue/geographic-tongue.
[24] Raviteja, "tongue2," *Kaggle*. [Online]. Available: https://www.kaggle.com/datasets/ravitejarj25/tongue2.
[25] 37, "舌像 Dataset," *Roboflow*, 2023. [Online]. Available: https://universe.roboflow.com/37/-tt3dc.
[26] C. Shorten, T. M. Khoshgoftaar, and B. Furht, *Text Data Augmentation for Deep Learning*, vol. 8, no. 1. Springer International Publishing, 2021.
[27] M. Temraz and M. T. Keane, "Solving the class imbalance problem using a counterfactual method for data augmentation," *Mach. Learn. with Appl.*, vol. 9, no. June, p. 100375, 2022.
[28] Jalu Nusantoro, Faldo Fajri Afrinanto, Wana Salam Labibah, Zamah Sari, and Yufis Azhar, "Detection of Covid-19 on X-Ray Image of Human Chest Using CNN and Transfer Learning," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 6, no. 3, pp. 430–441, 2022.
[29] D. Bhatt, C. Patel, H. Talsania, J. Patel, R. Vaghela, and S. Pandya, "Cnn7.Pdf," pp. 1–28, 2021.
[30] A. A. Reshi *et al.*, "An Efficient CNN Model for COVID-19 Disease Detection Based on X-Ray Image Classification," *Complexity*, vol. 2021, 2021.
[31] P. Sun *et al.*, "DL-IDS: Extracting features using CNN-LSTM hybrid network for intrusion detection system," *Secur. Commun. Networks*, vol. 2020, 2020.
[32] R. Yao, N. Wang, Z. Liu, P. Chen, and X. Sheng, "Intrusion Detection System in the Advanced Metering Infrastructure: A Cross-Layer Feature-Fusion CNN-LSTM-Based Approach," 2021.
[33] N. Ketkar and J. Moolayil, "Introduction to Machine Learning and Deep Learning," in *Deep Learning with Python*, 2021.
[34] H. Gholamalinezhad and H. Khosravi, "Pooling Methods in Deep Neural Networks, a Review," 2020.
[35] D. K. Baruah and K. Boruah, "Early Detection Of Canine Babesia From Red Blood Cell Images Using Deep Ensemble Learning," *J. Electron. Electromed. Eng. Med. Informatics*, vol. 6, no. 4, pp. 509–523, 2024.
[36] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, "A review of convolutional neural networks in computer vision," *Artif. Intell. Rev.*, vol. 57, no. 4, 2024.
[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–14, 2015.
[38] G. Sriram, T. R. G. Babu, R. Praveena, and J. V. Anand, "Classification of Leukemia and Leukemoid Using VGG-16 Convolutional Neural Network Architecture," *MCB Mol. Cell. Biomech.*, vol. 19, no. 1, pp. 29–40, 2022.
[39] S. M. Pradeep Singh, M. Shariff, D. P. Subramanyam, M. H. Varun, K. Shruthi, and A. S. Poornima, "Real Time Oral Cavity Detection Leading to Oral Cancer using CNN," *2023 Int. Conf. Network, Multimed. Inf. Technol. NMITCON 2023*, 2023.
[40] I. Konovalenko, P. Maruschak, J. Brezinová, J. Viňáš, and J. Brezina, "Steel surface defect classification using deep residual neural network," *Metals (Basel).*, vol. 10, no. 6, pp. 1–15, 2020.
[41] A. Victor Ikechukwu, S. Murali, R. Deepu, and R. C. Shivamurthy, "ResNet-50 vs VGG-19 vs training from scratch: A comparative analysis of the segmentation and classification of Pneumonia from chest X-ray images," *Glob. Transitions Proc.*, vol. 2, no. 2, pp. 375–381, 2021.
[42] N. Izdihar, S. B. Rahayu, and K. Venkatesan, "Comparison Analysis of CXR Images in Detecting Pneumonia Using VGG16 and ResNet50 Convolution Neural Network Model," *Int. J. Informatics Vis.*, vol. 8, no. 1, pp. 326–332, 2024.
[43] B. Mandal, A. Okeukwu, and Y. Theis, "Masked Face Recognition using ResNet-50," 2021.
[44] M. Vakili, M. Ghamsari, and M. Rezaei, "Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification," 2020.
[45] A. Churcher *et al.*, "An experimental analysis of attack classification using machine learning in IoT networks," *Sensors (Switzerland)*, vol. 21, no. 2, pp. 1–32, 2021.
[46] A. Agarwal, P. Sharma, M. Alshehri, A. A. Mohamed, and O. Alfarraj, "Classification model for accuracy and intrusion detection using machine learning approach," *PeerJ Comput. Sci.*, vol. 7, pp. 1–22, 2021.
[47] I. AKSAKALLI, S. KAÇDIOĞLU, and Y. S. HANAY, "Kidney X-ray Images Classification using Machine Learning and Deep Learning Methods," *Balk. J. Electr. Comput. Eng.*, vol. 9, no. 2, pp. 144–151, 2021.

AUTHOR BIOGRAPHY

Yufis Azhar received his Bachelor's degree in Computer Science from the Informatics Engineering program at Universitas Muhammadiyah Malang in 2009, and his Master's degree in Computer Science from Institut Teknologi Sepuluh Nopember, Surabaya, in 2013. Currently, he is a lecturer at the Informatics Study Program at Universitas Muhammadiyah Malang. His research interests include computer vision and machine learning. He has published numerous papers in reputable international journals and conferences. He also serves as a reviewer for several indexed journals and is actively involved in academic collaborations. In addition to his academic work, he is involved in various research projects focused on the application of AI technologies in real-world problems.

Fauzan Adrivano Setiono is from Blitar, East Java, previously attended SMK Islam 1 Blitar majoring in Computer Network Engineering. After that, he continued his education at Universitas Muhammadiyah Malang since 2020 and obtained a Bachelor of Computer Science degree from the Informatics Engineering study program at Universitas Muhammadiyah Malang in 2024. His research interests include data analytics and machine learning. He has a deep interest in this field and has participated in AI competitions by making AI-based applications to detect oral health.

Didih Rizki Chandranegara received his Bachelor's degree in Computer Science from the Informatics Department at Universitas Muhammadiyah Malang in 2014 and his Master's degree in Computer Science from Institut Teknologi Sepuluh Nopember, Surabaya, in 2018. He is currently a lecturer at the Informatics Department at Universitas Muhammadiyah Malang. His primary research interests lie in computer vision and machine learning, with several publications in prominent international journals and conferences. He also serves as a reviewer for various indexed journals and is actively engaged in academic collaborations and research

projects focused on applying artificial intelligence technologies to practical, real-world challenges.