Classification of Throat Disease Using CNNs: EfficientNetB0 and ResNet50

Aliyah Fajriyani*, I Gede Pasek Suta Wijaya, Fitri Bimantoro
Informatics Engineering Dept, Faculty of Engineering, University of Mataram
Jl. Majapahit 62, Mataram, Lombok NTB, INDONESIA
Email: fajriyanialiyah@gmail.com, [bimo, gpsutawijaya]@unram.ac.id

***Corresponding Author**

Abstract **Throat diseases are one of the global health issues. Early diagnosis could be an effective solution to prevent more severe throat disease. Automatic diagnosis based on medical images is possible to obtain by using Convolutional Neural Networks (CNN). This study employs two pretrained models namely ResNet50 and EfficientNetB0. The dataset contained 79 throat images divided to seven classes (normal, chronic laryngitis, acute pharyngitis, chronic pharyngitis, acute tonsillitis, chronic tonsillitis, and acute tonsillopharyngitis). The study was conducted in several scenarios and implemented gradually. First scenario, seven classes were merged into four classes (normal, pharyngitis, tonsillitis, and acute tonsillopharyngitis). Second scenario, four classes were combined into three classes (normal, pharyngitis, and tonsillitis). Third scenario, three classes were grouped into two classes (normal and illness). The results indicated that both the ResNet50 and EfficientNetB0 architectures achieved the highest performance in the third scenario (two classes). Both models showed identical evaluation matrics with accuracy of 91,67%, precision of 90%, recall of 100%, and F1-score of 94,74%. Furthermore, this study suggests that a dataset with numerous classes and limited data can be addressed by merging classes, thereby increasing the data size within each class.**

Key words: Classification, Throat Disease, CNN, ResNet50, EfficientNetB0.

I. Introduction

The throat plays a vital role in the human body, particularly in daily functions such as swallowing, breathing, and speaking. Due to its essential function and frequent use, the throat is highly susceptible to disorders caused by viral and bacterial infections, allergies, and environmental factors. Diseases of the upper respiratory tract, including throat infections, laryngitis, and nasopharyngeal cancer, continue to be significant global health issues. It is estimated that more than one billion cases of throat infections occur globally each year, with the highest prevalence in developing countries [1]. In 2022, the Indonesian Ministry of Health reported that diphtheria, one of the throat-related diseases, had spread to nearly all provinces in the country, including West Nusa Tenggara (NTB) [2].

The government has made various efforts to address this issue. In particular, the NTB provincial government has actively increased the provision of medical equipment to support diagnostic processes [3]. One of the primary tools used is the endoscope, which captures images that serve as the basis for physicians to make a diagnosis [4]. However, the current diagnostic process still relies heavily on manual examination by doctors, which can be time-consuming and prone to inaccuracies, especially in distinguishing between visually similar infections [5].

Furthermore, access to ENT specialists and diagnostic tools such as PCR or microbiological culture is very limited in rural areas and 3T regions (frontier, outermost, and disadvantaged areas) [5]. Early detection is crucial to prevent serious complications such as peritonsillar abscess or wider spread of infection [5]. To date, there is still no universal diagnostic method that is both fast and accurate in detecting various pathogens that cause throat infections, even though this region often serves as an entry point for new or mutated pathogens [6]. Therefore, there is a strong need for an automated system capable of accelerating image analysis and improving diagnostic precision and accuracy.

Many studies have been conducted on automated throat disease diagnosis systems. Previous study has explored expert systems based on certainty factor [7], Dempster-Shafer theory [8], and case-based reasoning [9] to detect throat cancer. However, these approaches were generally based only on patient-reported symptoms and did not utilize visual examination data, resulting in less accurate diagnoses [10]. The use of medical imaging for throat disease diagnosis can be enhanced through the application of deep learning, which enables automatic and more accurate disease classification. A prominent branch of deep learning is the Convolutional Neural Network (CNN) [10].

The reliability of Convolutional Neural Networks (CNNs) has been well established not only in the medical domain but also in other fields characterized by limited data availability. For instance, in the classification of local fruits in West Nusa Tenggara, CNN models such as ResNet50 and MobileNetV2 delivered high accuracy even with small datasets, especially when combined with preprocessing techniques like HSV color space transformation and background removal [11]. Similarly, in waste classification applications, models including ResNet50 and VGG16 achieved accuracy levels above 95%, with ResNet50 additionally demonstrating shorter training times in low-data environments [12]. These findings underscore the adaptability of CNNs to handle constrained and imbalanced datasets effectively, reinforcing their potential for use in image-based medical diagnosis systems.

Among the various available architectures, EfficientNetB0 and ResNet50 are frequently used due to their respective advantages. EfficientNetB0 is known for its efficiency in terms of model complexity and accuracy by incorporating compound scaling that balances depth, width, and input resolution, making it particularly suitable for small datasets while maintaining high performance [13]. In contrast, ResNet50 offers strong training stability and generalization ability through residual connections, which help overcome the vanishing gradient problem in deep networks [14].

In a previous study that compared EfficientNetB0, ResNet50, and MobileNetV3 in classifying acute pharyngitis using 343 throat images, EfficientNetB0 achieved the highest accuracy of 95.5%, followed by ResNet50 at 88.1% and MobileNetV3 at 82.1% [15]. Another study involving 339 throat images compared ResNet50, InceptionV3, and MobileNetV2 and found that ResNet50 achieved the best accuracy at 95.3% [16]. These findings demonstrate the significant potential of EfficientNetB0 and ResNet50 in accurately detecting throat diseases, making them promising candidates for further exploration in this study.

Based on this background, the present study aims to develop a throat disease classification model using endoscopic images by comparing the performance of two CNN architectures, namely EfficientNetB0 and ResNet50. To address limited data and class imbalance, the study applies basic image augmentation and class merging strategies based on anatomical and visual similarity. Additionally, Bayesian Optimization is used for efficient hyperparameter tuning to optimize model performance. This study is expected to provide insights into the most effective model for medical applications and to contribute to the development of faster, more accurate, and resource-efficient automated diagnostic systems, especially in healthcare settings with limited computational resources.

II. Literature Review

Study in throat disease classification based on endoscopic images has been widely conducted using transfer learning approaches involving Convolutional Neural Network (CNN) architectures. Numerous studies have demonstrated that CNNs can achieve high performance in analyzing throat images. For example, a study by Chng et al. in 2024 compared three CNN architectures, namely EfficientNetB0, ResNet50, and MobileNetV3, for the detection of acute pharyngitis using 343 throat images. The results showed that EfficientNetB0 achieved the highest accuracy at 95.5%, followed by ResNet50 with 88.1% and MobileNetV3 with 82.1% [15]. Similarly, Yoo et al. in 2020 employed ResNet50 enhanced with CycleGAN for data augmentation and achieved an accuracy of 95.3% [16]. In the context of laryngeal cancer detection, Xu et al. in 2023 used DenseNet201 on 2,254 laryngoscopy images and achieved a validation accuracy of 92% [17]. Another study by He et al. in 2021 implemented InceptionV3 to classify NBI and histopathological images and reported an AUC of 0.994 [18]. Furthermore, Alrowais et al. and Mohamed et al., both in 2023, applied hybrid architectures combining InceptionV3 with Aquila Optimization and EfficientNetB0 with Dwarf Mongoose Optimization for throat cancer classification. Their results achieved accuracies of 96.02% [19] and 99.53% [4] respectively. Based on the comparison of these studies, EfficientNetB0 and ResNet50 have consistently demonstrated superiority in both accuracy and parameter efficiency, which makes them appropriate choices for further exploration in this study.

Data limitations and class imbalance remain major challenges in medical image classification, as they can lead to overfitting and lower accuracy for minority classes [20]. One study proposed a strategy known as Class Confusion Merging, which aims to improve model accuracy by merging classes that are frequently misclassified based on the confusion matrix [21]. Although the method relies on the confusion matrix, the present study adopts a similar principle by grouping classes based on visual similarity and anatomical proximity. This approach is effective in reducing class imbalance within a small-scale endoscopic dataset [21]. In addition, basic image augmentation is used to increase data diversity without distorting essential features [22], as successfully implemented in previous study [15]. Compared to other techniques such as Synthetic Minority Oversampling Technique (SMOTE) or GAN-based augmentation, which may produce unrealistic images or require heavy computation, basic augmentation is more suitable for medical data that are sensitive to distortion [22].

For model optimization, Bayesian Optimization is widely used for model tuning due to its efficiency in exploring hyperparameter spaces. It has been shown to accelerate the tuning process while producing more stable and accurate CNN models, such as in brain tumor classification tasks [23]. This approach has also demonstrated superior performance in detecting ear diseases from otoscopic images, achieving accuracy as high as 98.10%, outperforming traditional manual tuning methods [24]. Compared to metaheuristic algorithms like Aquila Optimization and Dwarf Mongoose Optimization, which tend to be complex and less practical to implement, Bayesian Optimization offers a simpler yet effective alternative [25].

Considering all these approaches, this study presents a new contribution by implementing a combination of EfficientNetB0 and ResNet50 architectures, a class merging strategy based on visual and anatomical similarity, basic augmentation that is stable and efficient, and hyperparameter tuning using Bayesian Optimization. This study is specifically designed to perform optimally on small datasets and can be applied in healthcare facilities with limited computational resources. The novelty of this approach is expected to enrich the literature on automated throat disease diagnosis using deep learning and provide a practical solution for medical decision support systems.

III. Methodology

A. Study Flow

This study involved several stages, including data collection, data splitting, data preprocessing, data augmentation, and training models using ResNet50 and EfficientNetB0 architectures. The trained models were evaluated to assess their performance and determine whether it was optimal. If the performance was found to be optimal, the next step was to record and analyze the results. However, if the model's performance was not yet optimal, the training process was repeated by applying hyperparameter tuning using the Bayesian Optimization method. If the performance remained suboptimal even after tuning, the training process was repeated until optimal model performance was achieved by merging class. The overall study flow is illustrated in Figure 1.

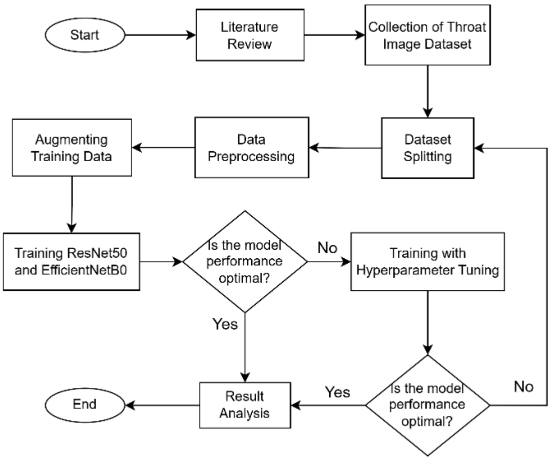
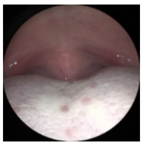
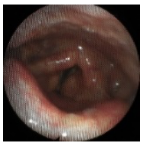
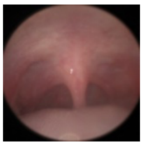
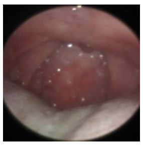
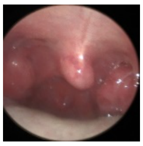
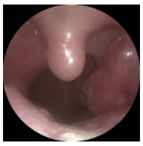
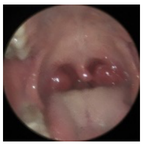

Start
Literature Review
Collection of Throat Image Dataset
Augmenting Training Data
Data Preprocessing
Dataset Splitting
Training ResNet50 and EfficientNetB0
Is the model performance optimal?
No
Training with Hyperparameter Tuning
Yes
End
Result Analysis
Yes
Is the model performance optimal?
No

Fig. 1. Study Flow.

B. Dataset Throat Diseases

This study used a dataset that consisting throat images captured by a laryngoscope. The dataset contains 79 throat images in 7 classes. The dataset was obtained from doctor practicing at University of Mataram Hospital. The dataset labeling process was conducted by two ENT specialists from the Faculty of Medicine, University of Mataram (UNRAM), namely Prof. Dr. dr. Hamsu Kadriyan, Sp.THT-KL (K), M.Kes, and Dr. dr. Didit Yudhanto, Sp.THT-KL, M.Sc. The distribution and sample of the dataset are presented in Table 1.

TABLE I. Distribution and Sample of The Dataset

Type of Disease	Total	Sample
Normal	20	
Chronic Laryngitis	3	
Acute Pharyngitis	16	
Chronic Pharyngitis	4	
Acute Tonsillitis	6	
Chronic Tonsillitis	18	
Acute Tonsillopharyngitis	12	

The dataset was divided into three sections, 70% of training data, 15% of validation data, and 15% of testing data. Model was trained by using training data and validation data. Meanwhile, model performance was evaluated using testing data.

C. Data Preprocessing

This stage aims to enhance image quality by applying resizing and rescaling techniques. Resizing was performed by changing the image dimensions from 512×512 pixels to 224×224 pixels to conform to the standard input size required by EfficientNetB0 [26] and ResNet50 [10]. Rescaling is intended to accelerate the training process and maintain model stability. For EfficientNetB0, pixel values were scaled from the original range of 0–255 to a range of 0 to 1. In contrast, for the ResNet50 model, pixel values were scaled from 0–255 to a range of -1 to 1.

D. Class Merging Startegy

To address the challenges of limited data and class imbalance, this study applied a step-by-step class merging strategy across four scenarios. The original dataset consisted of seven classes, which were progressively grouped based on visual similarity and anatomical proximity. In the first scenario, the dataset was reduced to four classes by merging acute and chronic forms of pharyngitis and tonsillitis. The second scenario further combined tonsillopharyngitis with tonsillitis, resulting in three classes. In the final scenario, all disease classes were merged into a single illness class, leading to binary classification between normal and illness. This merging process aimed to simplify the classification task while improving class distribution and data availability.

E. Data Augmentation

This process is carried out to increase the size and diversity of the dataset, address the imbalance class size, and reduce the risk of overfitting caused by the small dataset size. Augmentation was applied to the training data after the dataset had been divided into three subsets: training, validation, and testing data. The augmentation process was conducted using *Keras*'s *ImageDataGenerator*, involving horizontal flipping, width and height translation by 5%, random rotation between -10° and +10°, and zooming up to 20%. During augmentation, there were empty areas around the image because of image transformations such as rotation or translation. To resolve this, these empty areas were filled with colors interpolated from nearby pixels. This approach ensures that the augmented images remain natural and visually complete. It allowed the model to learn effectively without being distracted by missing or distorted parts. The number of augmented images in each class was determined as three times the number of images in the class with the highest original count. This total was then applied uniformly across all classes.

F. Model Architecture

In this process, the model was developed using a transfer learning approach due to the small dataset size, which can lead to overfitting and suboptimal model performance [13]. This study also employed pretrained models, EfficientNetB0 and ResNet50, for the classification task of throat diseases. The initial weights for training were obtained from ImageNet, followed by adjustments to the fully connected layers to accommodate the throat disease classification task. In this study, the architectures of ResNet50 and EfficientNetB0 were modified to support throat disease classification. This modification involved removing the original fully connected layers from each model and adding several new layers tailored to the target classes. The first step was to add a Global Average Pooling 2D layer to simplify the extracted features into a more compact form. The output was then passed through a dense layer with 128 neurons and a ReLU activation function, helping the model recognize important patterns in the data. Subsequently, a dropout layer with a rate of 0.5 was added to prevent overfitting. Finally, a dense output layer was added, adjusted to match the number of target classes.

G. Hyperparameter Tuning

Hyperparameter tuning can be performed manually by testing a predefined set of hyperparameters one by one. In this study, hyperparameter tuning was carried out automatically using the Bayesian Optimization method. This method was chosen for its ability to efficiently optimize the objective function by leveraging information from previous searches to determine the most promising next combination [23]. The range of values explored during the hyperparameter tuning process is presented in Table II.

TABLE II. Hyperparameter search range

Hyperparameter	**Value Range**
Unit Dense	32 to 512
Dropout Rate	0,2 to 0,5
Learning Rate	1e-6 to 1e-2

H. Model Evaluation

This process was conducted to evaluate the performance of the ResNet50 and EfficientNetB0 models in classifying throat diseases. The evaluation was carried out by testing the models on the test dataset using four metrics: accuracy, precision, recall, and F1-score [27]. All metrics were calculated based on the values of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) derived from the confusion matrix. The formulas for each metric based on the confusion matrix are presented below.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

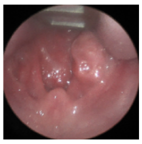
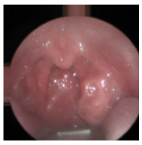
$$F1 - score = 2 \times \frac{(precision \times recall)}{(precision + recall)} \quad (4)$$

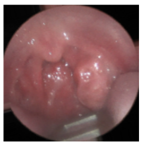
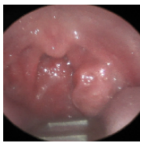
IV. Result and Discussion

A. Data Augmentation

Augmentation was applied only to the training data to increase variation and address class imbalance within the dataset. The augmentation techniques used include flipping, translation, rotation, and zooming. The results of the augmentation process are presented in Table III.

TABLE III. Augmentation Techniques

Augmentation	**Before**	**After**
Flip		
Translation		

Augmentation	Before	After
Rotation		
Zoom		

After the augmentation process was applied to the training data, the number of augmented samples for each scenario of dataset division can be seen in Table IV.

TABLE IV. AUGMENTATION DATA

Scenario	Class	Total	
		Before	After
Dataset of Seven Classes	Normal	20	60
	Chronic Laryngitis	3	60
	Acute Pharyngitis	16	60
	Chronic Pharyngitis	4	60
	Acute Tonsillitis	6	60
	Chronic Tonsillitis	18	60
	Acute Tonsillopharyngitis	12	60
Dataset of Four Classes	Normal	20	72
	Pharyngitis	23	72
	Tonsillitis	24	72
	Acute Tonsillopharyngitis	12	72
Dataset of Three Classes	Normal	20	108
	Pharyngitis	23	108
	Tonsillitis	36	108
Dataset of Two Classes	Normal	20	177
	Illness	59	177

B. Model Evaluation

The performance of the trained ResNet50 and EfficientNetB0 models was evaluated for the task of throat disease classification. The training process was conducted over 100 epochs with a learning rate of 1e-4 (0.0001), utilizing the Adam optimizer. To enhance training stability, callbacks were applied, including EarlyStopping which halts the training process when the *val_loss* metric stops improving and ModelCheckpoint, which stores the highest *val_accuracy* value achieved during training [13]. The performance of both ResNet50 and EfficientNetB0 models was assessed using a confusion matrix and standard evaluation metrics, namely accuracy, precision, recall, and F1-score. The following sections present the evaluation results of the models under several different experimental scenarios.

B.1. Dataset of Seven Classes

In this scenario, the dataset used is the original dataset consisting of seven classes. The test data from this dataset was evaluated using both the ResNet50 and EfficientNetB0 models. The confusion matrix resulting from the evaluation of the ResNet50 dan EfficientNetB0 model is presented in Figure 2 dan Figure 3.

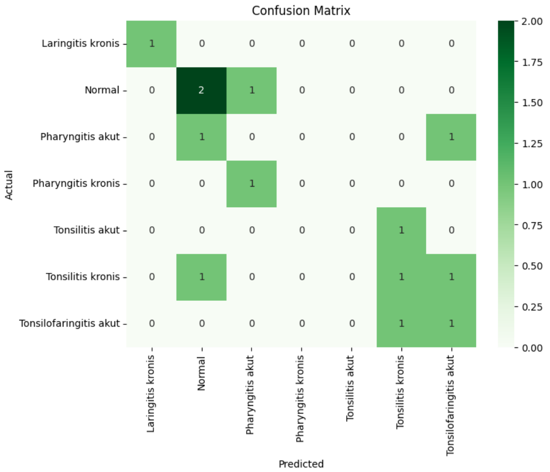


Fig. 2. Confusion Matrix ResNet50.

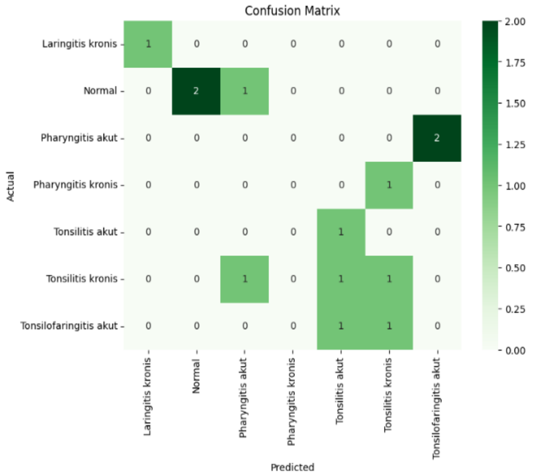


Fig. 3. Confusion Matrix EfficientNetB0.

Based on Figures 2 and 3, both models still encountered difficulties in distinguishing images between classes. Both ResNet50 and EfficientNetB0 frequently misclassified acute tonsillitis and chronic tonsillitis. Specifically, the ResNet50 model tended to predict images of acute tonsillitis as chronic tonsillitis, while EfficientNetB0 often predicted chronic tonsillitis as acute tonsillitis. Additionally, ResNet50 misclassified images of chronic pharyngitis as acute pharyngitis. This misclassification may occur because the differentiation between acute and chronic conditions is primarily based on the duration of the illness experienced by the patient, whereas their anatomical locations are the same and their visual characteristics are nearly identical [28].

Both models also misclassified normal images as acute pharyngitis. On the other hand, ResNet50 incorrectly classified images of acute pharyngitis and chronic tonsillitis as normal. In contrast, EfficientNetB0 did not misclassify any diseased images as normal. Although both models successfully predicted chronic laryngitis images correctly, this result may be biased due to the small class size and extensive data augmentation [29]. The models' difficulty in distinguishing between classes is further supported by the evaluation metric values presented in Table V.

TABLE V. PERFORMANCE EVALUATION OF THE SEVEN-CLASS SCENARIO

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
ResNet50	38,46	30,95	35,71	32,93
EfficientNetB0	38,46	38,10	42,86	37,62

Table V shows that both models achieved relatively low accuracy, precision, recall, and F1-score values, ranging from approximately 30% to 40% [27]. In the context of medical technology, recall indicates the model's ability to correctly identify patients who are actually ill, whereas precision reflects the model's ability to avoid misclassifying healthy individuals as diseased [30]. The low evaluation metric values suggest that neither model was able to accurately detect diseases or predict each class reliably.

In this scenario, both models experienced overfitting. The models adapted too closely to the training data and failed to generalize well to the test data [13]. Overfitting may have been caused by a small and imbalanced dataset [29]. As a corrective strategy, a subsequent scenario was implemented involving class merging to reduce the models' classification difficulty and enhance their performance in throat disease classification. Class merging under specific conditions can be an effective solution when facing challenges related to limited dataset size [31].

B.2. Dataset of Four Classes

In this scenario, the original seven-class dataset was restructured into four classes: normal, pharyngitis, tonsillitis, and tonsillopharyngitis. The merging of acute and chronic cases of the same disease was carried out by considering their anatomical location and the visual similarity of the images [28]. The chronic laryngitis class was merged into the laryngitis class, as the anatomical location can visually overlap, particularly when inflammation spreads [28]. The normal and acute tonsillopharyngitis classes were retained due to their distinct visual characteristics and the sufficient number of available samples [29]. The confusion matrices resulting from the evaluation of the ResNet50 and EfficientNetB0 models in this scenario are presented in Figures 4 and 5.

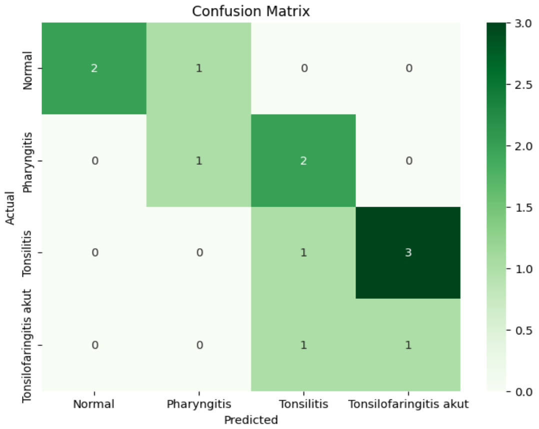


Fig. 4. Confusion Matrix ResNet50.

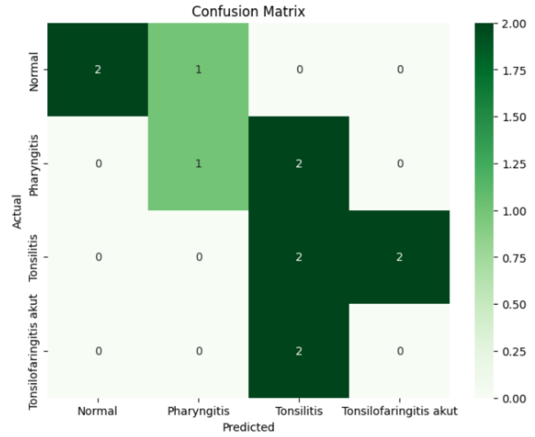


Fig. 5. Confusion Matrix EfficientNetB0.

Based on Figure 4, the ResNet50 model no longer misclassified diseased images as normal. However, both models still misclassified one normal image as acute pharyngitis. In addition, both models incorrectly classified pharyngitis images as tonsillitis. The models also continued to confuse tonsillitis with acute tonsillopharyngitis, and vice versa. This confusion may have occurred because tonsillopharyngitis is a combined condition involving both tonsillitis and pharyngitis [28]. When the inflammation is more prominent in the tonsils, tonsillopharyngitis images are likely to be interpreted by the model as tonsillitis.

These findings indicate that both models still struggle to differentiate between throat disease classes. This limitation may be due to the insufficient number of images per class, which hinders the models' ability to fully learn and recognize throat disease patterns [29]. Additionally, low-quality or unclear medical images can present further challenges for classification models [32]. The difficulty faced by the models in distinguishing between classes is further supported by the evaluation metrics shown in Table VI.

TABLE VI. PERFORMANCE EVALUATION OF THE FOUR-CLASS SCENARIO

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
ResNet50	41,67	50	43,57	44,58
EfficientNetB0	41.67	45,83	37,50	40

Table VI indicates that both models experienced a slight performance improvement in this scenario compared to the previous scenario with seven classes. Despite the improvement, the overall performance of both models remains suboptimal [27]. This is evident from the evaluation metric values, which are still relatively low, ranging between 40% and 50%. The fact that precision scores are higher than recall scores may also suggest that both models are experiencing overfitting [13]. This issue is primarily attributed to the limited amount of data and the imbalance in class distribution [29].

In this scenario, Hyperparameter tuning was applied to improve model performance by searching for the best hyperparameters. The hyperparameter tuning was conducted using the Bayesian Optimization method. After completing the tuning process using the specified method, the best hyperparameters obtained are presented in Table VII.

TABLE VII. BEST PARAMETERS IN THE FOUR-CLASS SCENARIO

Model	Hyperparameter	Value Range
ResNet50	Unit Dense	384
	Dropout Rate	0,4
	Learning Rate	0.00011964743859134101
EfficientNetB0	Unit Dense	256
	Dropout Rate	0,2
	Learning Rate	0.0006748735068204596

The comparison of two models with hyperparameter tuning can be seen in Table VIII.

TABLE VIII. COMPARISON TWO MODELS WITH HYPERPARAMETER TUNING IN THE FOUR-CLASS SCENARIO

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
BO ResNet50	83,3	66,7	66,7	66,7
BO EfficientNetB0	79,2	58,3	58,3	58,3

After tuning, the performance of both the ResNet50 and EfficientNetB0 models improved, with ResNet50 slightly outperforming EfficientNetB0. This suggests that ResNet50 is more sensitive to hyperparameter configurations, whereas EfficientNetB0 tends to be more stable [33]. Nevertheless, the performance of both models remains suboptimal despite the application of hyperparameter tuning. Therefore, a subsequent scenario will be conducted.

B.3. Dataset of Three Classes

In the previous scenario, both models frequently confused tonsillitis with tonsillopharyngitis. Considering this, the tonsillopharyngitis class was merged into the tonsillitis class. As a result, this scenario includes three classes: normal, pharyngitis, and tonsillitis. After testing with the ResNet50 and EfficientNetB0 models, identical confusion matrices were obtained. The confusion matrices for both models are presented in Figure 6.

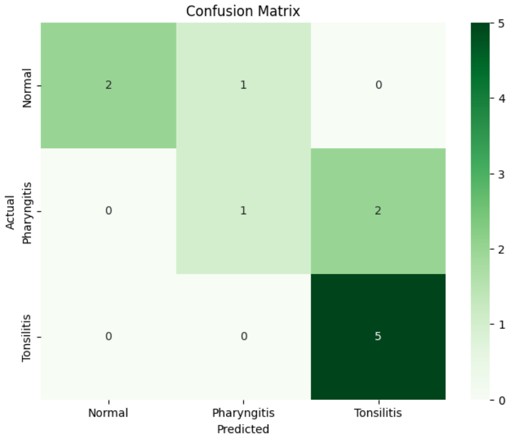


Fig. 6. Confusion Matrix ResNet50 dan EfficientNetB0.

Based on Figure 6, in this scenario, both models successfully predicted all tonsillitis images accurately. However, both models consistently misclassified one normal image as pharyngitis and two pharyngitis images as tonsillitis. The models still tended to interpret pharyngitis images as tonsillitis. This may occur because, during a pharyngitis episode, the tonsils can also become inflamed [28]. A pharyngitis image may be misclassified as tonsillitis if the inflammation appears more prominent in the tonsillar region. The corresponding evaluation metric values are presented in Table IX.

TABLE IX. PERFORMANCE EVALUATION OF THE THREE-CLASS SCENARIO

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
ResNet50	72,73	73,81	66,67	67,78
EfficientNetB0	72,73	73,81	66,67	67,78

Table IX shows that the ResNet50 and EfficientNetB0 models achieved improved accuracy, precision, recall, and F1-score values, ranging from approximately 60% to 70%. The identical results between the two models suggest that both made correct and incorrect predictions at the same points. This may be attributed to the small size of the test dataset, which could lead the models to recognize similar patterns with limited variation [29]. Interestingly, in the three-class classification scenario, the model performed better without data augmentation. This may be due to the high visual similarity between pharyngitis and tonsillitis, where basic augmentation techniques such as rotation or flipping could obscure important distinguishing features between the classes. Instead of improving performance, augmentation on a small dataset may introduce variations that are not clinically relevant, potentially reducing the model's accuracy. This indicates that data augmentation does not always enhance model performance, especially when the added variations fail to reflect meaningful or discriminative patterns relevant to the target classes. To further enhance model performance, hyperparameter tuning using Bayesian Optimization was applied in this scenario. The best hyperparameters obtained are presented in Table X.

TABLE X. BEST PARAMETERS IN THE THREE-CLASS SCENARIO

Model	Hyperparameter	Value Range
ResNet50	Unit Dense	174
	Dropout Rate	0.4855450747417324
	Learning Rate	0.0004015275311354817
EfficientNetB0	Unit Dense	92
	Dropout Rate	0.4077502503662843
	Learning Rate	0.0020145932338176123

The comparison of two models with hyperparameter tuning can be seen in Table XI.

TABLE XI. COMPARISON TWO MODELS WITH HYPERPARAMETER TUNING IN THE THREE-CLASS SCENARIO

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
BO ResNet50	81,82	90,48	77,78	77,78
BO EfficientNetB0	72,73	73,81	66,67	67,78

After tuning, the ResNet50 model demonstrated improved performance. In contrast, the performance of EfficientNetB0 did not show any improvement. This difference in response to tuning indicates that ResNet50 is more flexible with respect to configuration adjustments, allowing the tuning process to enhance its ability to recognize disease patterns [33]. On the other hand, EfficientNetB0, which is designed with an efficient architecture, tends to be stable but less responsive to hyperparameter changes, particularly when applied to small datasets [33].

B.4. Dataset of Two Classes

In this scenario, the pharyngitis and tonsillitis classes were merged into a single "diseased" class, while the normal class was retained. This merging was conducted because, in the previous scenario, both models struggled to distinguish between pharyngitis and tonsillitis. After testing with the ResNet50 and EfficientNetB0 models, identical confusion matrices were obtained. The confusion matrices for both models are presented in Figure 7.

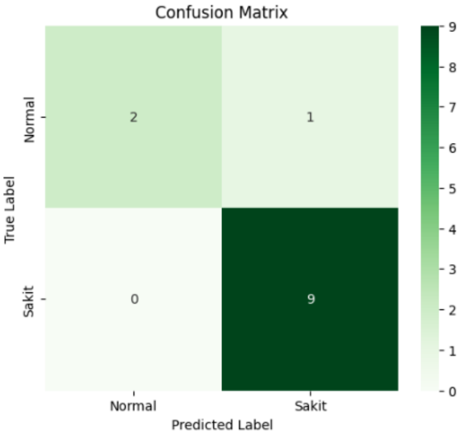


Fig. 7. Confusion Matrix of ResNet50 and EfficientNetB0.

Based on Figure 7, both models correctly predicted all images in the diseased class. However, both models consistently misclassified one normal image as diseased. When the two previously separate classes were merged, the models succeeded in making correct predictions. Nevertheless, misclassification in the class that was not merged still occurred. The evaluation metric values are presented in Table XII.

TABLE XII. PERFORMANCE EVALUATION OF THE TWO-CLASS SCENARIO

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
ResNet50	91,67	90	100	94,74
EfficientNetB0	91,67	90	100	94,74

The perfect recall value in Table XII indicates that the model did not miss a single diseased case in the test data, which is crucial in a medical context, as false negatives can pose serious risks [30]. However, the gap between precision and recall suggests the presence of false positives, specifically one normal case misclassified as diseased. Moreover, both models were still unable to overcome overfitting despite the implementation of EarlyStopping. In addition, the use of data augmentation in this two-class classification did not improve model performance, which contrasts with the results observed in the three-class classification. This is likely because the classification task is simpler, making the models less dependent on additional data variation. Therefore, further improvements are necessary to enhance the model's ability to accurately identify the normal class, in order to prevent healthy individuals from being misdiagnosed.

Overall, this scenario demonstrates improved performance compared to the previous multi-class scenarios. The class merging strategy successfully enhanced the model's predictive capability [33]. However, this high performance may not fully reflect the model's capacity in more complex classification tasks. Such simplification may obscure important distinctions between diseases and still leaves the model vulnerable to overfitting due to the limited dataset size and class imbalance.

V. ConClusion and Recommendations

Based on the findings of this study, it can be concluded that progressive class merging on a small dataset contributes to improved performance of CNN models in the task of throat disease classification. The accuracy, precision, recall, and F1-score of the two model architectures used ResNet50 and EfficientNetB0 increased as the number of classes was gradually reduced. Importantly, the class merging was carried out under specific considerations. In the case of throat diseases, the merging was based on anatomical location and visual similarity in medical images. Additionally, the ResNet50 model responded more positively to hyperparameter tuning than EfficientNetB0, which tended to be more stable.

This study has several limitations. First, the dataset size was very small (only 79 images), which may lead to overfitting and reduce the model's ability to generalize to new data. Second, the imbalance in data distribution across classes required heavy augmentation in the smallest classes, which could affect classification accuracy, particularly in the multi-class scenario. Third, although binary classification yielded higher performance scores, it does not necessarily indicate that the model is capable of detecting specific disease types, as the classification was limited to distinguishing between normal and diseased cases without specifying the disease type in detail.

For further model development, it is recommended to increase the amount of original data for each class, particularly for classes with very small sample sizes. This will allow the model to learn more representatively and reduce dependence on augmented data. Additionally, exploring other architectures is advisable to identify more optimal models for the multi-class classification of throat diseases. Incorporating model interpretability features, such as Grad-CAM, is also important to enable visualization of important image regions and support medical validation processes. Moreover, the findings of this study can be further developed into web-based or mobile applications to assist in the automatic early detection of throat diseases. Lastly, the use of ensemble methods can also be considered to combine the strengths of multiple model architectures in order to improve accuracy and robustness of the classification system.

References

[1] Miller K, Carapetis J, Van Beneden C, Cadarette D, Daw J, Moore H, Bloom D, and Cannon J, "The global burden of sore throat and group A Streptococcus pharyngitis: A systematic review and meta-analysis," *EClinicalMedicine*, vol. 48, p. 101458, 2022, doi: 10.1016/j.

[2] Kementrian Kesehatan, *Profil Kesehatan Indonesia*. 2024.

[3] RSUD Provinsi NTB, "Reviu II Rencana Strategis (RENSTRA) Tahun 2019-2023," 2020.

[4] Mohamed N, Almutairi R, Abdelrahim S, Alharbi R, Alhomayani F, Elamin Elnaim B, Elhag A, and Dhakal R, "Automated Laryngeal Cancer Detection and Classification Using Dwarf Mongoose Optimization Algorithm with Deep Learning," *Cancers (Basel)*, vol. 16, no. 1, Jan. 2024, doi: 10.3390/cancers16010181.

[5] L. Lukama, C. Aldous, C. Michelo, and C. Kalinda, "Ear, Nose and Throat (ENT) disease diagnostic error in low-resource health care: Observations from a hospital-based cross-sectional study," *PLoS One*, vol. 18, no. 2 February, Feb. 2023, doi: 10.1371/journal.pone.0281686.

[6] I. Senapati, A. Pradhan, L. Senapati, R. Prasath, and T. Swarnkar, "VADDOT: Vocal Analysis to Detect Diseases of Throat," in *2023 2nd International Conference on Ambient Intelligence in Health Care (ICAIHC)*, 2023, pp. 1–6. doi: 10.1109/ICAIHC59020.2023.10431445.

[7] M. Permata Sari and A. Desiani, "Diagnosa Penyakit THT (Telinga, Hidung, Tenggorokan) Menggunakan Metode Certainty Factor pada Sistem Pakar," *J. Artif. Intell. Softw. Eng.*, vol. 3, no. 1, p. 7, 2023, doi: 10.30811/jaise.v3i1.3902.

[8] M. Dwi Lestari Harianja, Ishak, and S. Yakub, "Implementasi Sistem Pakar untuk Mendiagnosa Penyakit Faringitis (Radang Tenggorokan) Menggunakan Metode Dempster Shafer," *Jurnal Sistem Informasi TGD*, vol. 3, no. 5, pp. 773–781, Sep. 2024, [Online]. Available: https://ojs.trigunadharma.ac.id/index.php/jsi

[9] P. Tarigan, "Sistem Pakar Metode Case Based Reasoning Mendiagnosa Penyakit Kanker Tenggorokan," *KAKIFIKOM (Kumpulan Artikel Karya Ilmiah Fakultas Ilmua Komputer)*, vol. 04, no. 02, pp. 100–107, Oct. 2022, [Online]. Available: http://www.ejournal.ust.ac.id/index.php/KAKIFIKOM/article/view/2444%0Ahttp://www.ejournal.ust.ac.id/index.php/KAKIFIKOM/article/view/2444/2103

[10] N. P. Maylianti, G. Ngurah, L. Wijayakusuma, P. Chandra, and A. Wiguna, "Comparison of EfficientNet-B0 and ResNet-50 for Detecting Diseases in Cocoa Fruit," *Journal of Applied Informatics and Computing (JAIC)*, vol. 9, no. 1, pp. 115–120, Feb. 2025, [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[11] M. Azzam Al Husaini, A. Yudo Husodo, and F. Bimantoro, "Classification of Local Fruit Types using Convolutional Neural Network Method (Study Case: Lombok Island)." [Online]. Available: http://jcosine.if.unram.ac.id/

[12] P. Y. Andrean, F. Bimantoro, and R. P. Rassy, "Comparative Analysis of ResNet-50 and VGG16 Architecture Accuracy in Garbage Classification System." [Online]. Available: http://jcosine.if.unram.ac.id/

[13] M. Alruwaili and M. Mohamed, "An Integrated Deep Learning Model with EfficientNet and ResNet for Accurate Multi-Class Skin Disease Classification," *Diagnostics*, vol. 15, no. 5, Mar. 2025, doi: 10.3390/diagnostics15050551.

[14] T. Shahzad, T. Mazhar, S. M. Saqib, and K. Ouahada, "Transformer-inspired training principles based breast cancer prediction: combining EfficientNetB0 and ResNet50," *Sci Rep*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-98523-w.

[15] S. Y. Chng, P. J. W. Tern, M. R. X. Kan, and L. T. E. Cheng, "Deep Learning Model and its Application for the Diagnosis of Exudative Pharyngitis," *Healthc Inform Res*, vol. 30, no. 1, pp. 42–48, Jan. 2024, doi: 10.4258/hir.2024.30.1.42.

[16] T. K. Yoo, J. Y. Choi, Y. Jang, E. Oh, and I. H. Ryu, "Toward Automated Severe Pharyngitis Detection with Smartphone Camera Using Deep Learning Networks," *Comput Biol Med*, vol. 125, Oct. 2020, doi: 10.1016/j.compbiomed.2020.103980.

[17] Z. H. Xu, D. G. Fan, J. Q. Huang, J. W. Wang, Y. Wang, and Y. Z. Li, "Computer-Aided Diagnosis of Laryngeal Cancer Based on Deep Learning with Laryngoscopic Images," *Diagnostics*, vol. 13, Dec. 2023, doi: 10.3390/diagnostics13243669.

[18] He Y, Cheng Y, Huang Z, Xu WHu R, Cheng L, He S, Yue C, Qin G, Wang Y, and Zhong Q, "A Deep Convolutional Neural Network-Based Method for Laryngeal Squamous Cell Carcinoma Diagnosis," *Ann Transl Med*, vol. 9, no. 24, Dec. 2021, doi: 10.21037/atm-21-6458.

[19] F. Alrowais, K. Mahmood, S. S. Alotaibi, M. A. Hamza, R. Marzouk, and A. Mohamed, "Laryngeal Cancer Detection and Classification Using Aquila Optimization Algorithm with Deep Learning on Throat Region Images," *IEEE Access*, vol. 11, pp. 115306–115315, 2023, doi: 10.1109/ACCESS.2023.3324880.

[20] C. J. Hellín, A. A. Olmedo, A. Valledor, J. Gómez, M. López-Benítez, and A. Tayebi, "Unraveling the Impact of Class Imbalance on Deep-Learning Models for Medical Image Classification," *Applied Sciences (Switzerland)*, vol. 14, no. 8, Apr. 2024, doi: 10.3390/app14083419.

[21] X. Miao, Y. Zhang, J. Zhang, and X. Liang, "Hyperspectral Image Classification Based on Class Confusion Merging and Soft Band Selection," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2021, pp. 3645–3648. doi: 10.1109/IGARSS47720.2021.9553136.

[22] E. Goceri, "Medical image data augmentation: techniques, comparisons and interpretations," *Artif Intell Rev*, vol. 56, no. 11, pp. 12561–12605, 2023, doi: 10.1007/s10462-023-10453-z.

[23] Nawabi J, Eminovic S, Hartenstein A, Baumgaertner G, Schnurbusch N, Rudolph M, Wasilewski D, Onken J, Siebert E, Wiener E, Bohner G, Dell'Orco A, Wattjes M, Hamm B, Fehrenbach U, and Penzkofer T, "Bayesian-Optimized Convolutional Neural Networks for Classifying Primary Tumor Origin of Brain Metastases from MRI," *Brain Sci*, vol. 15, no. 5, p. 450, Apr. 2025, doi: 10.3390/brainsci15050450.

[24] H. M. Afify, K. K. Mohammed, and A. E. Hassanien, "Insight into Automatic Image Diagnosis of Ear Conditions Based on Optimized Deep Learning Approach," *Ann Biomed Eng*, vol. 52, no. 4, pp. 865–876, Apr. 2024, doi: 10.1007/s10439-023-03422-8.

[25] J. O. Agushaka, A. E. Ezugwu, and L. Abualigah, "Dwarf Mongoose Optimization Algorithm," *Comput Methods Appl Mech Eng*, vol. 391, p. 114570, 2022, doi: https://doi.org/10.1016/j.cma.2022.114570.

[26] Nur Fajrina A, Hanni Pradana Z, Indah Purnama S, Romadhona S, Studi P, Telekomunikasi T, and Biomedis T, "Application of the EfficientNet-B0 Architecture in the Classification of Acute Lymphoblastic Leukemia," *Jurnal Riset Rekayasa Elektro*, vol. 6, no. 1, pp. 59–68, Jun. 2024, [Online]. Available: http://jurnalnasional.ump.ac.id/index.php/JRRE

[27] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, vol. 17, no. 1, pp. 168–192, 2018, doi: 10.1016/j.aci.2018.08.003.

[28] M. Mohammad and M. Suhail, *Textbook of Ear, Nose and Throat Diseases, 11th Edition*.

[29] N. De La Fuente, M. Majó, I. Luzko, H. Córdova, G. Fernández-Esparrach, and J. Bernal, "Enhancing Image Classification in Small and Unbalanced Datasets through Synthetic Data Augmentation," Sep. 2024, [Online]. Available: http://arxiv.org/abs/2409.10286

[30] Hicks S, Strümke I, Thambawita V, Hammou M, Riegler M, Halvorsen P, and Parasa S, "On evaluation metrics for medical applications of artificial intelligence," *Sci Rep*, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-022-09954-8.

[31] P. Georgiadis, E. V. Gkouvrikos, E. Vrochidou, T. Kalampokas, and G. A. Papakostas, "Building Better Deep Learning Models Through Dataset Fusion: A Case Study in Skin Cancer Classification with Hyperdatasets," *Diagnostics*, vol. 15, no. 3, Feb. 2025, doi: 10.3390/diagnostics15030352.

[32] Ma J, Nakarmi U, Kin C, Sandino C, Cheng J, Syed A, Wei P, Pauly J, andVasanawala S, "Diagnostic Image Quality Assessment and Classification in Medical Imaging: Opportunities and Challenges," in *Proceedings - International Symposium on Biomedical Imaging*, IEEE Computer Society, Apr. 2020, pp. 337–340. doi: 10.1109/ISBI45749.2020.9098735.

[33] C. Coppola, L. Papa, M. Boresta, I. Amerini, and L. Palagi, "Tuning parameters of deep neural network training algorithms pays off: a computational study," *TOP*, Oct. 2024, doi: 10.1007/s11750-024-00683-x.