Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics e-ISSN: 2656-8624
Homepage: https://ijeeemi.org/; Vol. 8, No. 3, pp. 299-316, August 2026

RESEARCH PAPER

OPEN ACCES

Data Splitting Strategies for Down Syndrome Facial Classification: A Comparative Study Using EfficientNet-B0 and MobileNetV2

Dzaky Dhiya Ul-Haq[1], Yunidar Yunidar[1], Melinda Melinda[1], Nurlida Basir[2], and Rosmawinda[1]

[1] Department of Electrical Engineering and Computer, Universitas Syiah Kuala, Banda Aceh, Indonesia
[2] Faculty of Science and Technology, Universiti Sains Islam Malaysia (USIM), Nilai Negeri Sembilan, Malaysia

Abstract

Early identification of Down Syndrome (DS) is essential for timely intervention; however, conventional diagnostic approaches often require specialized clinical expertise and significant resources. Recent advances in deep learning-based facial image analysis offer a promising alternative, yet the impact of data partitioning strategies on model performance and stability remains insufficiently explored. This study investigates the effects of different data-splitting strategies on DS facial image classification using EfficientNet-B0. A total of 3,030 facial images were collected from Roboflow and curated through preprocessing techniques, including Gaussian noise reduction, image sharpening, and contrast enhancement. Two data partitioning configurations, 70:20:10 and 80:10:10, were evaluated using five-fold cross-validation. Model performance was assessed using accuracy, precision, recall, and F1-score, while statistical significance was examined using the Friedman test. The results show that the 70:20:10 configuration achieved an average accuracy of 87.88% ± 3.03%, while the 80:10:10 configuration achieved a slightly higher accuracy of 89.09% ± 2.53%. The Friedman test indicates statistically significant differences ($p < 0.05$). However, the improvement is relatively marginal, with a small-to-moderate effect size (Cohen's $d = 0.43$) and no significant difference in variance ($p > 0.05$), indicating limited practical significance. A trade-off between accuracy and evaluation stability was observed. While the 80:10:10 configuration benefits from a larger training set, the 70:20:10 configuration provides more stable and balanced performance, particularly in minimizing false negatives. These findings highlight that higher accuracy does not necessarily imply more reliable or clinically meaningful performance, emphasizing the importance of appropriate data partitioning in medical image classification.

Paper History

Received April 08, 2026
Revised May 25, 2026
Accepted May 30, 2026
Published June 02, 2026

Keywords

Down Syndrome;
EfficientNet-B0;
Friedman Test;
Cross-Validation;
Deep Learning

Author Email

dzaky.d@mhs.usk.ac.id
yunidar@usk.ac.id
melinda@usk.ac.id
nurlida@usim.edu.my
rosmawinda@mhs.usk.ac.id

I. Introduction

Down Syndrome (DS), also referred to as trisomy 21, is recognized as the most prevalent genetic disorder globally, resulting from the existence of an extra copy of chromosome 21. Individuals diagnosed with DS display unique craniofacial features that can serve as visual biomarkers for automated analysis [1] [2]. While DS can be clinically diagnosed at birth, the issue of delayed diagnosis persists as a significant concern, as it hampers access to early intervention programs crucial for enhancing cognitive development, functional independence, and long-term quality of life [3], [4], [5]. In developing nations like Indonesia, the detection of DS faces additional obstacles due to the scarcity of genetic specialists, disparities in healthcare access, and inadequate diagnostic infrastructure, which frequently leads to late or erroneous diagnoses [4]. Traditional diagnostic methods predominantly depend on clinical observation and invasive prenatal techniques, which, although reliable for diagnosis, carry medical risks, substantial financial burdens, and logistical challenges that hinder the implementation of large-scale screening [6]. These issues highlight the pressing necessity for non-invasive, accessible, and cost-effective postnatal screening solutions for DS [7].

Deep learning, particularly convolutional neural networks (CNNs), has demonstrated strong capability in medical image analysis by automatically learning discriminative features from raw data [8], [9], [10]. In facial-based DS detection [11], [12], CNN models have shown promising performance in identifying morphological patterns associated with DS, enabling automated screening systems. Furthermore, lightweight architectures such as MobileNetV2 have been widely adopted due to their efficiency and suitability for limited computational environments, while still maintaining

Corresponding author: Yunidar, yunidar@usk.ac.id, Department of Electrical Engineering and Computer, Universitas Syiah Kuala, Banda Aceh, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.338

competitive classification performance [13], [14]. However, most existing studies primarily focus on model architecture improvements, with limited attention given to experimental design factors that may affect the reliability and validity of the results.

Despite the promising performance of deep learning-based DS detection models, their reliability and generalizability are strongly influenced by experimental design choices, especially dataset partitioning strategies. Data splitting determines how available samples are allocated into training, validation, and testing subsets, directly affecting both model learning capacity and evaluation robustness [15], [16]. Several studies have demonstrated that different train-test split ratios significantly affect model performance and stability, and that increasing the amount of training data does not always guarantee better generalization, potentially introducing variability in results [17]. Moreover, improper data splitting strategies may lead to biased evaluation or unrealistic performance estimates, especially when random splits are used without accounting for data distribution characteristics [18].

Dataset splitting is a fundamental in machine learning. It ensures models are tested on unseen data. Typically, datasets are divided into training, validation, and test subsets. The training set teaches model parameters. The validation set tunes the model. The test set evaluates generalization. The split ratio balances learning and reliable evaluation. A poor split can cause underfitting, overfitting, or bias [17]. Thus, proper data partitioning is key to stable, generalizable results. To address these limitations, this study develops an automated DS detection system based on facial images using two deep learning architectures: EfficientNet-B0 and MobileNetV2. EfficientNet-B0, designed with compound scaling to network depth, width, and resolution, enables high accuracy at relatively low computational cost [18], [19]. In contrast, MobileNetV2 employs depthwise separable convolutions and inverted residual blocks, delivering a lightweight and efficient solution for real-world applications with limited computational resources [13], [20]. This study comprehensively evaluates how each model performs under different data-splitting strategies to clarify the strengths and limitations of both architectures in DS detection.

The primary objective of this research is not only to analyze the impact of different dataset partitioning strategies on model performance, but also to provide a structured evaluation framework for understanding how data splitting influences learning stability and generalization in deep learning-based DS classification. Specifically, this study compares two commonly used data splitting configurations, namely 70:20:10 and 80:10:10, for training, validation, and testing sets. Beyond direct comparison, this works aims to explore how variations in training data proportion affect model robustness across different architectures, namely EfficientNet-B0 and MobileNetV2. Furthermore, this study seeks to establish an initial foundation for adaptive data splitting strategies by identifying performance trends and trade-offs across different partitioning schemes. Rather than assuming a fixed optimal ratio, the findings of this study are intended to support the development of more data-driven and context-aware splitting strategies in future research. To ensure robust and unbiased evaluation, a 5-fold cross-validation scheme is applied [21], and statistical significance is assessed using the Friedman test [22]. By combining empirical evaluation with statistical validation, this work aims to contribute not only to model performance analysis but also to the methodological design of reliable and generalizable deep learning experiments.

II. Methods

A. System Workflow

This study adopts a structured workflow to evaluate the performance of two deep learning models: EfficientNet-B0 and MobileNetV2. EfficientNet-B0 and MobileNetV2 are convolutional neural network architectures commonly used for image classification tasks. The workflow compares different data-splitting strategies, which determine how the dataset is divided into training and validation sets. It is designed for fair, systematic, and reproducible comparisons between the proposed data partitioning configurations. The process also maintains robust model validation procedures. The complete process includes dataset acquisition (gathering raw data), preprocessing (preparing data for use), data partitioning (dividing data into subsets), model training, cross-validation (assessing model performance across different splits), and statistical evaluation (analyzing results quantitatively).

A facial image dataset is collected and labeled into two classes: Down Syndrome (DS) and non-Down Syndrome (Non-DS). Before model training, all images are preprocessed to improve data quality and ensure feature consistency. Preprocessing includes resizing (changing image size to a standard shape), normalization (scaling pixel values to a standard range), image enhancement (adjusting contrast, brightness, or sharpness), and augmentation (creating a modified version using rotations, flips, or lighting changes) to improve robustness to variations in lighting, pose, and image quality. After preprocessing, the dataset is split into training, validation, and test sets with 70:20:10 and 80:10:10 ratios. These approaches test how varying proportions of training data affect model performance and evaluation stability.

For each data splitting strategy, a five-fold cross-validation (5-fold CV) scheme is applied to the training set: the training set is divided into five equally-sized subsets, each serving as a validation set once. This improves generalization capability and reduces bias caused by random data partitioning. Both EfficientNet-B0 and MobileNetV2 are trained independently on each fold. Validation performance is monitored to assess learning behaviour across folds. Performance metrics from each fold are recorded for later analysis. The classification results from all folds and data-splitting configurations are statistically evaluated. The Friedman test, a

Corresponding author: Yunidar, yunidar@usk.ac.id, Department of Electrical Engineering and Computer, Universitas Syiah Kuala, Banda Aceh, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.338

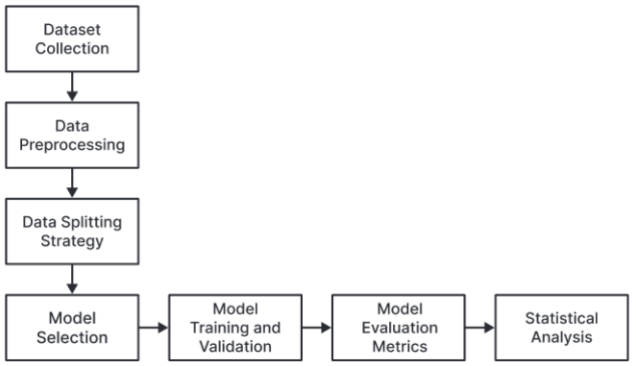


Fig. 2. Proposed deep learning workflow for facial classification using EfficientNet-B0 and MobileNetV2 under different data-splitting.

nonparametric statistical method that does not assume data come from a specific distribution, is used to compare multiple related samples. This analysis determines whether performance differences among models and data-splitting strategies are statistically significant. The complete system workflow is illustrated in Fig. 1.

B. Dataset Collection

The facial image dataset used in this study was obtained from the Roboflow platform [23], which provides a publicly available dataset for computer vision research. Initially, a total of 6,332 facial images were collected and labeled into two classes: namely Down Syndrome (DS) and non-Down Syndrome (Non-DS). Examples of facial images from both classes are shown in Fig. 2.

(a) (b)

Fig. 1. Sample facial images showing (a) Down Syndrome and (b) non-Down Syndrome

Before model development, a data screening process ensured image quality and uniformity. Images with insufficient resolution, significant blur, occlusion, or improper facial alignment were excluded. This initial filtering reduced the dataset to 3,030 images. A second, more stringent quality control stage was then applied to further refine the dataset by removing ambiguous, duplicated, or borderline-quality samples. As a result, the dataset was reduced to a total of 2,620 images for experimentation. Further clarification of the dataset's composition is essential for ensuring transparency. The dataset comprises facial images of children ranging from 0 to 15 years, as indicated in the initial documentation. The images display differences in lighting, facial orientation, background intricacy, and overall quality. Comprehensive metadata, including ethnicity, geographic origin, or the condition under which the images were acquired (such as camera type and environment), is not readily accessible. This absence arises from the data being publicly aggregated. Such a limitation may lead to demographic bias and influence the generalizability of the model. To mitigate potential data-related issues, duplicate or visually identical images were removed during preprocessing. The curated dataset was then used for preprocessing, partitioning, and model training. Various data-splitting strategies were applied as detailed in the following sections. All data splitting used random shuffling with a fixed seed to ensure reproducibility.

The dataset was compiled from various sources that lacked explicit subject identifiers. Consequently, a rigorous strategy for splitting data by subject could not be applied. This situation presents a potential risk of data leakage, as multiple images of the same person might be present in the training, validation, and testing subsets. Although efforts were made to reduce duplication, complete elimination of identity-level overlap is not feasible due to the lack of subject-level annotations. This limitation may lead to somewhat optimistic performance estimates; therefore, the findings should be interpreted with caution. Nevertheless, the dataset still provides sufficient variability and diversity to support the evaluation of data-splitting strategies for deep learning-based DS facial classification. Despite the relatively small size of the dataset in this study, which includes 3,030 images after initial filtering and 2,620 images for final experimentation, efforts were made to ensure a balanced class distribution between DS and non-DS samples. This balance helps minimize classification bias and facilitates a fair evaluation of model performance across both classes. To compensate for the limited dataset size, data augmentation techniques-such as horizontal flipping, slight rotation, brightness and contrast adjustment, and minor zooming-were applied exclusively to the training dataset. These augmentations simulate real-world variations, including differences in lighting, pose, and scale, thereby improving model robustness and reducing the risk of overfitting. However, the relatively small dataset size remains a significant limitation of this study. Since deep learning models typically require larger, more diverse datasets to capture broader populations and real-world clinical scenarios. As a result, the scalability and generalizability of the proposed approach may be constrained. To address these limitations, future work should focus on expanding the dataset, incorporating cross-dataset validation, and including a more diverse range of demographic representations to enhance model reliability.

C. Data Processing

Data preprocessing was performed to enhance image quality and improve the robustness of the classification model [24], [25]. Facial images were obtained from the Roboflow platform. All filtering and quality control procedures were conducted before preprocessing. This yielded a refined dataset for subsequent experiments. A

Corresponding author: Yunidar, yunidar@usk.ac.id, Department of Electrical Engineering and Computer, Universitas Syiah Kuala, Banda Aceh, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.338

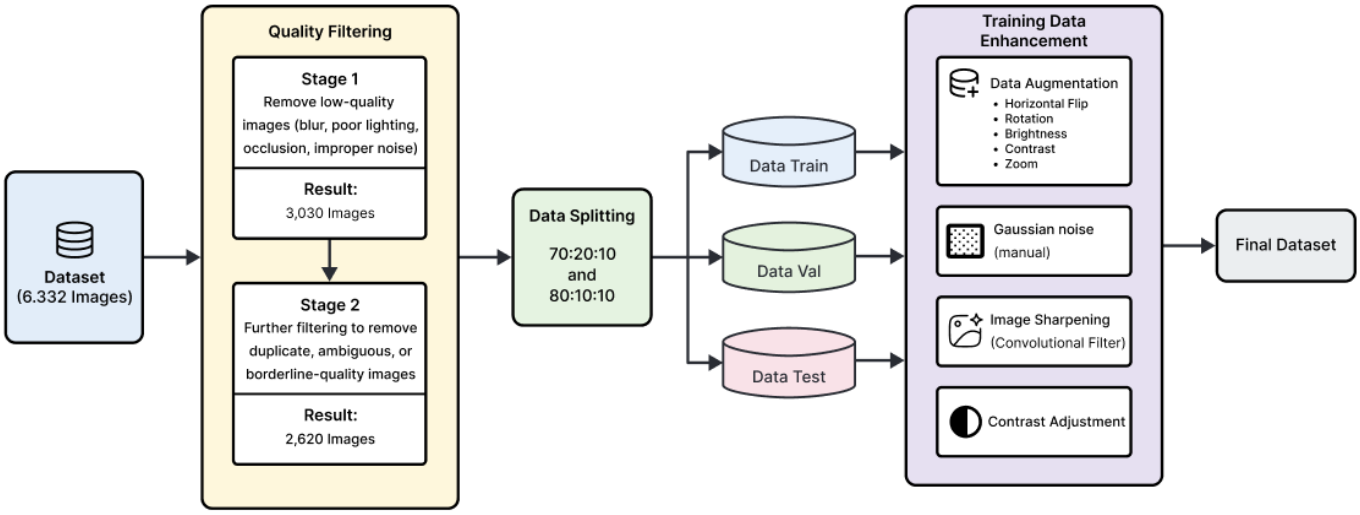


Fig. 3 Data preprocessing flowchart with resizing splitting and augmentation processes.

customized preprocessing and augmentation strategy was implemented on the training dataset. This method mimicked real-world variations while maintaining crucial facial features. Various augmentation techniques were utilized: horizontal flipping (50%), minor angle rotation (±10 °), brightness modification (±20%), contrast modification (±20%), and slight zooming (0.9-1.1).

These transformations added variability in pose, lighting, and scale without considerably changing facial structure. In addition to geometric and photometric augmentation, Gaussian noise was manually applied to a portion of the training data. This simulated sensor noise and real-world image imperfections [26], [27]. The approach aimed to improve the model's robustness against variations in image acquisition conditions. Unlike geometric transformations, Gaussian noise introduces subtle pixel-level perturbations that mimic realistic noise patterns. However, introducing artificial Gaussian noise during medical image preprocessing is not a standard practice and may potentially reduce clinical realism. Therefore, an ablation study was conducted to evaluate the impact of Gaussian noise by comparing model performance with and without noise augmentation. This analysis aims to assess whether including noise improves model robustness or introduces performance bias.

Image sharpening was applied using convolutional filters to enhance edge details and emphasize facial features [28] [29]. Contrast adjustment was also incorporated. This improved the distinction between light and dark regions, reduced the impact of illumination variations, and enhanced feature consistency across samples [30], [31]. All preprocessing and augmentation techniques were applied exclusively to the training dataset. The validation and test datasets remained unchanged to ensure unbiased performance evaluation. This strategy aimed to improve generalization, reduce overfitting, and support stable learning under different data-splitting configurations [32].

D. Data Splitting Strategy

To evaluate the impact of different data partitioning strategies on model performance, this study applies two commonly used data split configurations, namely 70:20:10 and 80:10:10, for training, validation, and testing sets. Both strategies were designed to maintain class balance between DS and Non-DS samples, ensuring a fair and unbiased comparison. The distribution of samples for the 70:20:10 data split configuration is presented in Table 1, showing the allocation of DS and Non-DS images across the training, validation, and testing subsets.

Following the initial split, each training set was further evaluated using five-fold cross-validation to improve generalization capability and reduce dependence on a single data partition. In the 70:20:10 configuration, the balanced allocation of training and validation data enables stable model tuning, whereas the 80:10:10 configuration emphasizes a larger training set to assess whether increased data exposure enhances learning. The corresponding sample distribution for the 80:10:10 configuration in Table 2, which provides the number of DS and Non-DS images assigned to each subset. The purpose of this evaluation is to determine if the impact of data splitting strategies remains consistent across two different deep learning architectures, EfficientNet-B0 and MobileNetV2. To ensure that any observed performance

Table. 1 Dataset distribution using 70:20:10 split for training validation testing subsets.

Subsets	Down Syndrome	Non-DS	Total
Training	916	916	1,832
Validation	262	262	524
Testing	132	132	264
Total	1,310	1,310	2,620

Corresponding author: Yunidar, yunidar@usk.ac.id, Department of Electrical Engineering and Computer, Universitas Syiah Kuala, Banda Aceh, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.338

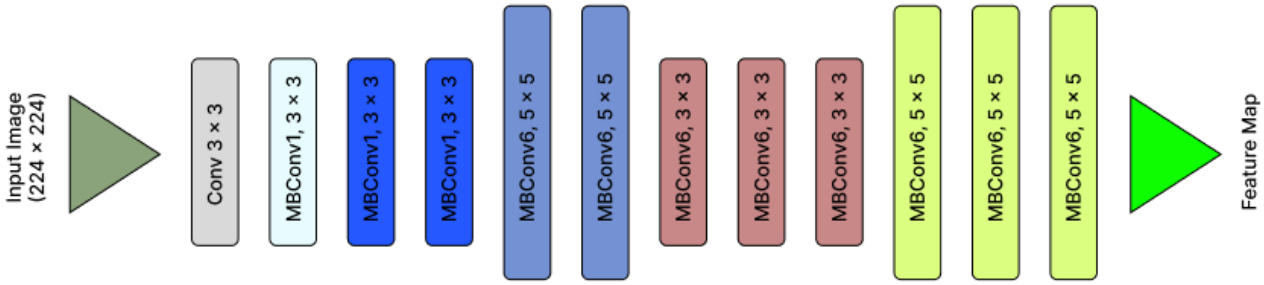


Fig. 4. EfficientNet-B0 architecture showing convolution blocks scaling and feature extraction process.

differences are due to the data splitting strategy rather than data imbalance or sample size variation, identical dataset sizes and class distributions are used for both configurations. The results are then statistically analyzed using the Friedman test to determine if performance differences across configurations and models are significant. Beyond direct comparison, this design also aims to provide insights into how data partitioning influences model stability and generalization, contributing toward the development of more adaptive and data-driven splitting strategies in deep learning experiments.

E. Model Architecture

This study uses EfficientNet-B0 and MobileNetV2 for facial image classification due to their strong performance and computational efficiency, achieved through network depth, width, and input resolution optimization. Compared to deeper models, EfficientNet-B0 balances capability and simplicity, making it ideal for experiments with different data splits [28], [33]. Fig. 4. shows the EfficientNet-B0 model architecture, while Fig. 5. presents the MobileNetV2 architecture.

The EfficientNet-B0 architecture consists of a series of MBConv blocks, combined with a Squeeze-and-Excitation (SE) mechanism, that enhances channel-wise feature recalibration and improves discriminative feature learning. The input images are resized to 224 x 224 pixels, following the standard EfficientNet-B0 configuration, and then passed through successive convolutional and pooling layers to extract hierarchical facial features [34]. MobileNetV2 is a streamlined convolutional neural network designed for efficient processing, especially in resource-constrained environments. It employs depthwise separable convolutions to separate spatial filtering from feature combination, thereby lowering both computational complexity and the number of parameters compared to traditional convolutions [35]. The architecture is built on inverted residual blocks: feature maps are enlarged using pointwise convolutions, spatial features are gathered through depthwise convolutions, and the output is mapped to a reduced-dimensional space [35]. Linear bottleneck layers further retain essential features during dimensionality reduction, boosting efficiency without compromising performance. This structure enables MobileNetV2 to achieve high accuracy while maintaining a low computational cost, making it ideal for real-time and mobile applications [36].

Table. 2 Dataset distribution using 80:10:10 split for training validation testing subsets.

Subsets	Down Syndrome	Non-DS	Total
Training	1,048	1,048	2,096
Validation	131	131	262
Testing	131	131	262
Total	1,310	1,310	2,620

For the classification task, the original fully connected layers of both EfficientNet-B0 and MobileNetV2 are replaced with a custom classification head. This head uses global average pooling to reduce spatial dimensions, followed by a fully connected layer with ReLU activation and a dropout layer to mitigate overfitting. The final output layer uses a sigmoid activation function for binary classification between DS and Non-DS classes. All models are initialized with ImageNet-pretrained weights for faster convergence and transfer learning. During training, the base architectures are fine-tuned to better adapt to facial features for DS classification. The same configurations-architecture, initialization, and training procedure-is applied across both data split strategies (70:20:10 and 80:10:10) for fair comparison.

F. Training Model and Validation

EfficientNet-B0 and MobileNetV2 were deployed for binary classification between DS and Non-DS facial images. Training used a binary cross-entropy loss function, suitable for two-class tasks. The Adam optimizer, set at a 0.0001 learning rate, balanced convergence speed and training stability [37], [38]. During training, input images were processed in mini-batches of 16. Each model was trained for 50 epochs per experimental configuration. A checkpoint mechanism automatically saved weights with the best validation performance, ensuring the best generalization. For each data splitting strategy (70:20:10 and 80:10:10), the training procedure was run independently using five-fold cross-validation. For each fold, the model was trained on part of the training data. The rest was used for validation. This reduces bias from a single data partition and provides more robust, reliable model evaluation.

Corresponding author: Yunidar, yunidar@usk.ac.id, Department of Electrical Engineering and Computer, Universitas Syiah Kuala, Banda Aceh, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.338

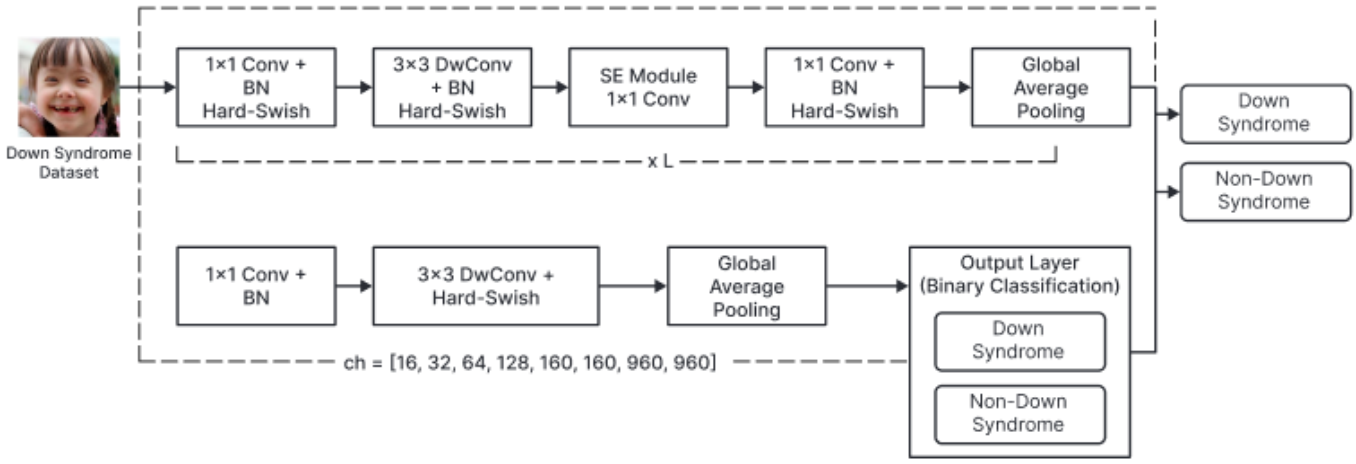


Fig. 5. MobileNetV2 architecture illustrating inverted residual blocks with linear bottlenecks for efficient feature extraction.

Table. 3. Model training hyperparameter including optimizer learning rate batch size epochs.

Parameter	Value
Architecture	EfficientNet-B0, MobileNetV2
Loss Function	Binary Cross-Entrophy
Optimizer	Adam
Learning Rate	0.0001
Batch Size	16
Epoch	50
Input Image Size	224 x 224
Cross-Validation	5-Fold
Model Selection	Best Validation Loss

All training metrics, including training and validation losses and accuracy, were recorded for subsequent analysis. The same training procedure and hyperparameter configuration were consistently applied to both EfficientNet-B0 and MobileNetV2 across all experimental setups to ensure a fair and controlled comparison. The hyperparameter configuration used in this study is summarized in Table 3. These settings were selected based on prior empirical studies and preliminary experiments to ensure stable training behavior and fair performance comparison. In particular, initial trials were conducted to identify suitable values for the learning rate, batch size, and number of epochs that provide consistent convergence across folds. Although an extensive hyperparameter optimization strategy, such as grid search or Bayesian optimization, was not employed due to computational constraints, the selected configuration produced stable, reliable performance. The same parameters were therefore fixed across all experiments to isolate the effect of dataset partitioning and preprocessing techniques on model performance. Future work may explore automated hyperparameter optimization to further improve model performance.

G. Model Evaluation Metrics

The performance of each classification model was evaluated using a confusion matrix, which provides a detailed representation of the model's ability to distinguish between DS and Non-DS classes. Based on the confusion matrix, four commonly used metrics in binary classification were computed, namely accuracy, precision, recall, and F1-score, as defined in Eq. (1)-(4) [39].

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

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

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

$$F1 - Score = 2\,x\,\frac{Precision\ x\ Recall}{Precision + Recall} \quad (4)$$

In Eq. (1)-(4), TP represents true positives, TN denotes true negatives, FP indicates false positives, and FN represents false negatives. To obtain a more reliable estimation of model performance, five-fold cross-validation (5-fold CV) was employed [40]. The dataset was divided into K=5 equally sized folds; in each iteration, one fold was used for validation, and the remaining folds for training. The evaluation process was repeated across all folds, and the average performance was computed as:

$$\bar{M} = \frac{1}{K}\sum_{k=1}^{K} M_k \quad (5)$$

In Eq. (5), $\bar{M}$ denotes the average validation performance, K represents the total number of folds in cross-validation, and M_k indicates the validation performance obtained from the k-th fold.

Corresponding author: Yunidar, yunidar@usk.ac.id, Department of Electrical Engineering and Computer, Universitas Syiah Kuala, Banda Aceh, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.338

To evaluate the statistical significance of performance differences across configurations, this study employs a nonparametric test, namely the Friedman test. This approach is selected because cross-validation folds are not independent, as they share portions of the training data, leading to correlated performance estimates and violating the independence assumption required by parametric statistical tests. The Friedman test evaluates differences among multiple models by ranking their performance across each fold [22]. For each fold, models are assigned ranks based on their performance, where the best-performing model receives rank 1. The average rank for each model is then computed, and the Friedman test statistic is defined as:

$$x_f^2 = \frac{12N}{k(k+1)}\left[\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4}\right] \quad (6)$$

In Eq. (6), x_f^2 represents the Friedman test statistic, N denotes the number of cross-validation folds, k is the number of models being compared, and $R_j$is the average rank of the j- th model. The test statistic approximately

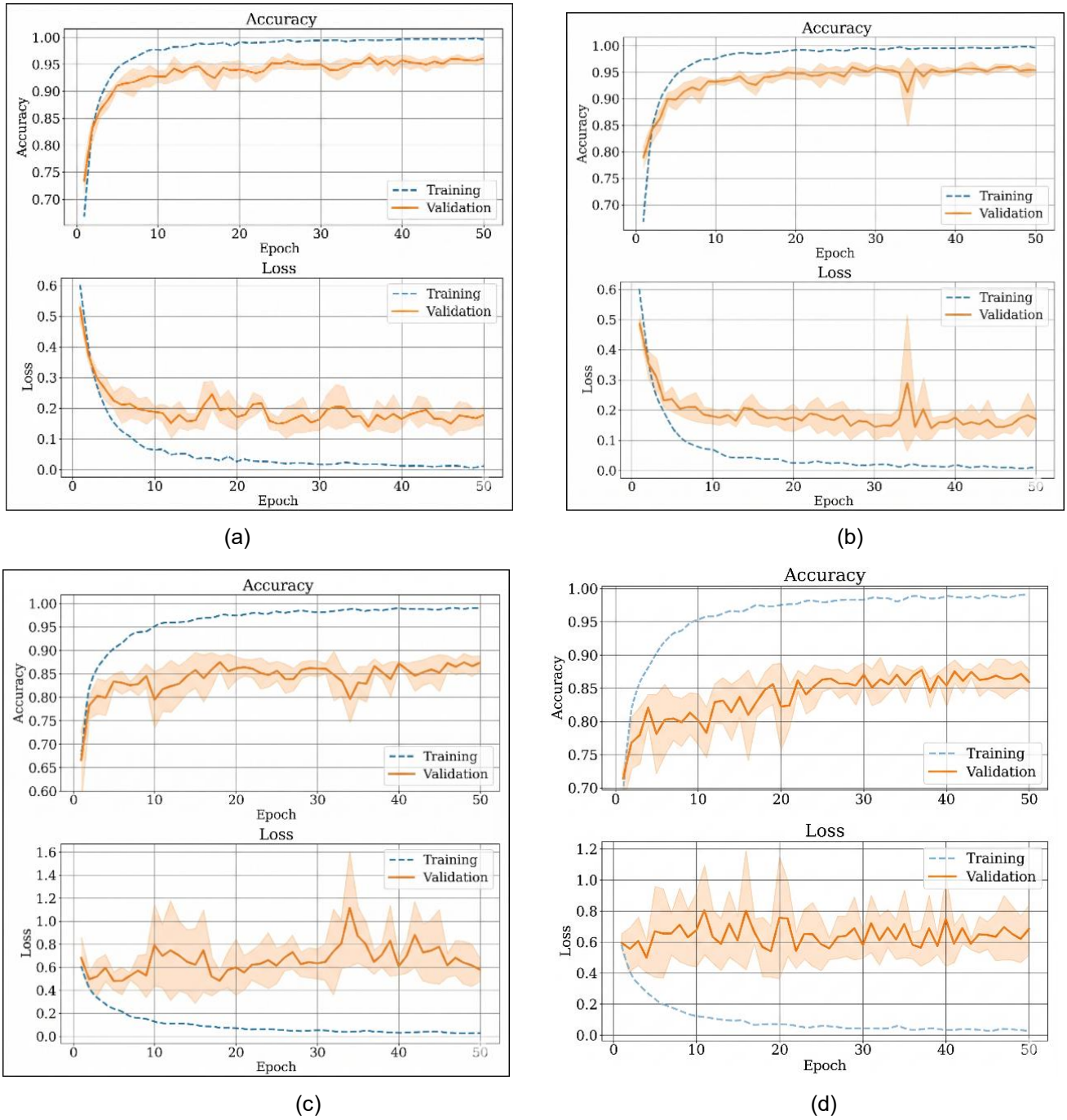


Fig. 6. Learning curves (accuracy and loss) of EfficientNet-B0 and MobileNetV2 across different data split configurations: (a) EfficientNet-B0 with 70:20:10 split, (b) EfficientNet-B0 with 80:10:10 split, (c) MobileNetV2 with 70:20:10 split, and (d) MobileNetV2 with 80:10:10 split.

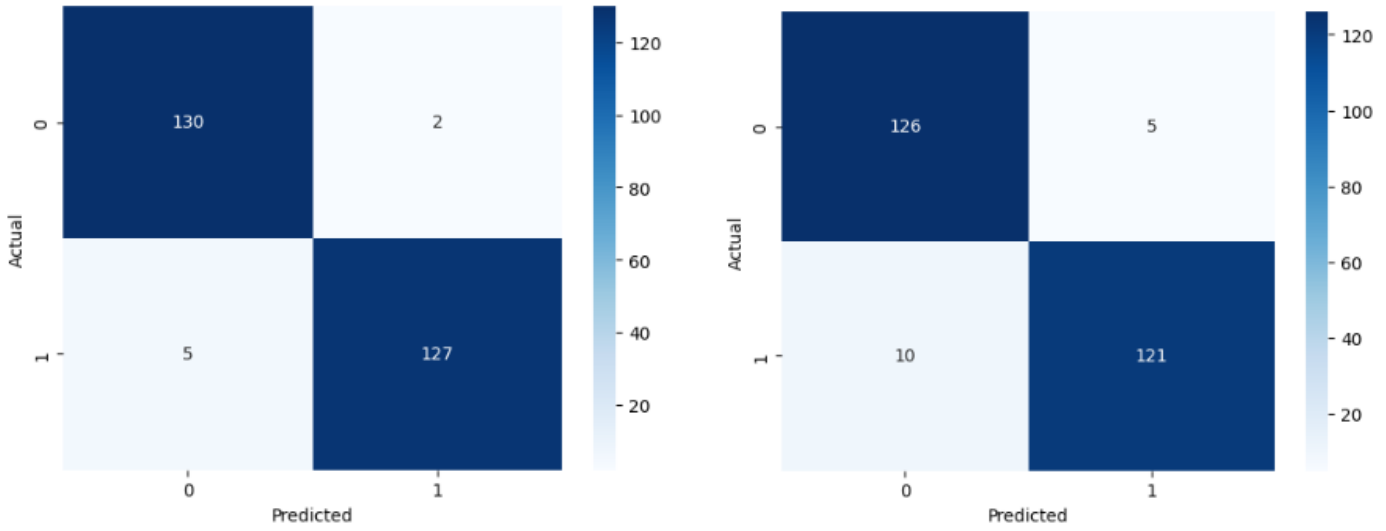
Corresponding author: Yunidar, yunidar@usk.ac.id, Department of Electrical Engineering and Computer, Universitas Syiah Kuala, Banda Aceh, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.338

follows a chi-square distribution with $k-1$ degrees of freedom. If the computed statistic exceeds the critical value or the p-value is less than the significance level ($\alpha = 0.05$), the null hypothesis indicating no significant difference among models, is rejected. In addition to statistical significance testing, the practical significance of performance differences was evaluated using effect size measurement. Cohen's d was used to quantify the magnitude of difference between two configurations, defined as:

$$d = \frac{M_1 - M_2}{SD_{pooled}} \quad (7)$$

Table. 4. Comparative performance of EfficientNet-B0 and MobileNetV2 under different data splits.

Model	Split	Precision	Recall	Specificity	F1-Score
EfficientNet-B0	70:20:10	0.9845	0.9621	0.9848	0.9732
EfficientNet-B0	80:10:10	0.9603	0.9237	0.9618	0.9416
MobileNetV2	70:20:10	0.9048	0.8636	0.9091	0.8837
MobileNetV2	80:10:10	0.7605	0.9695	0.6947	0.8523



(a) (b)

Fig. 7. Confusion matrices of EfficientNet-B0 under different data split configurations: (a) 70:20:10 and (b) 80:10:10.

In Eq. (7), M_1 and M_2 present the mean performance of the two configurations, and SD_{pooled} denotes the pooled standard deviation, which was calculated as:

$$SD_{pooled} = \sqrt{\frac{SD_1^2 + SD_2^2}{2}} \quad (8)$$

In Eq. (8), SD_1 and SD_2 denote the standard deviations of the two configurations being compared. This metric provides a standardized measure of effect size, enabling interpretation of whether observed differences are small, moderate, or large in practical terms. Confidence intervals were also considered to assess the stability and overlap of performance estimates across configurations. To further evaluate the stability of model performance, a variance comparison was conducted using Levene's test. This test assesses whether the variances of performance metrics across cross-validation folds differ significantly between configurations and is robust to deviations from normality, making it suitable for small-sample experimental settings.

H. Statistical Analysis

Statistical analysis was conducted to evaluate the significance of performance variations across different data splits using a 5-fold cross-validation scheme. In this approach, the dataset was divided into five subsets, with each fold used once for validation and the remaining folds for training. The average validation performance across folds was computed to obtain a robust estimation of model performance, as defined in Eq. (5). The average validation accuracy for each fold was calculated and then aggregated to produce an overall performance measure across all folds. This approach provides a comprehensive evaluation by capturing both fold-specific variability and overall model performance, thereby reducing dependency on a single data partition. To assess whether the observed performance differences across models and data-splitting strategies are statistically significant, a nonparametric Friedman test was employed. Unlike parametric methods

Corresponding author: Yunidar, yunidar@usk.ac.id, Department of Electrical Engineering and Computer, Universitas Syiah Kuala, Banda Aceh, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.338

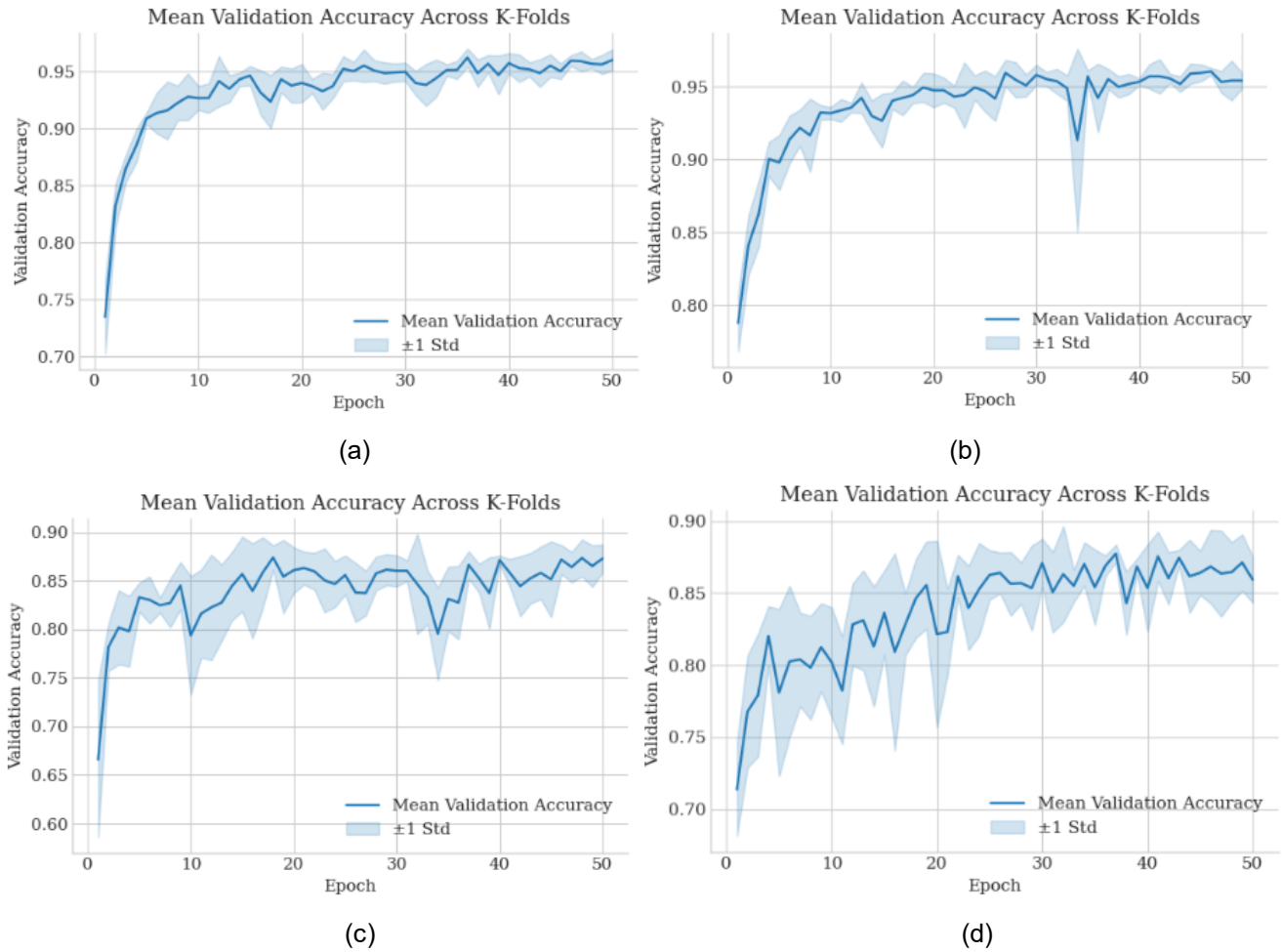


Fig. 8. M Mean validation accuracy across K-folds for EfficientNet-B0 and MobileNetV2 under different data split configurations: (a) EfficientNet-B0 with 70:20:10 split, (b) EfficientNet-B0 with 80:10:10 split, (c) MobileNetV2 with 70:20:10 split, and (d) MobileNetV2 with 80:10:10 split.

such as ANOVA, the Friedman test does not assume independence or normal distribution of the observations, making it more suitable for repeated cross-validation experiments where performance measurements are inherently correlated. The Friedman test ranks model performance across folds, with the best-performing model receiving the highest rank. The average rank of each model is then computed across all folds, and the Friedman test statistic is used to determine whether there are significant differences among the compared models. The null hypothesis (H_0) assumes that there is no statistically significant difference in model performance across the evaluated configurations, while the alternative hypothesis (H_1) indicates that at least one model performs significantly differently. A significance level of $p < 0.05$ was used to determine statistical significance. This statistical framework ensures that model evaluation is not dependent on a single data split, but instead reflects consistent performance across multiple folds, thereby improving the reliability and robustness of the experimental results.

III. Results

A. Training and Validation Performance Under Different Data Splits

To evaluate the effect of different data partitioning strategies on model learning behavior, EfficientNet-B0 and MobileNetV2 were trained using two split configurations, namely 70:20:10 and 80:10:10. The validation accuracy curves for all configurations are presented in Fig. 6(a)–(d). For EfficientNet-B0, both configurations achieved stable convergence with high validation performance across the five-fold cross-validation process. In the 70:20:10 configuration, the model achieved a mean validation accuracy of 95.98% with a standard deviation of 0.99%, corresponding to a variance value of 0.00010, as shown in Table 6. In comparison, the 80:10:10 configuration achieved a slightly lower mean validation accuracy of 95.80% with a higher standard deviation of 1.34% and variance of 0.00018. These results indicate that the 70:20:10 configuration provides more stable learning performance across folds. The learning curves in Fig. 6(a) and Fig. 6(b) further demonstrate that EfficientNet-B0 converges rapidly within the early training epochs and maintains relatively smooth validation trends throughout training. Although the 80:10:10 configuration visually appears

Corresponding author: Yunidar, yunidar@usk.ac.id, Department of Electrical Engineering and Computer, Universitas Syiah Kuala, Banda Aceh, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.338

Table. 5. Ablation study on preprocessing techniques using 70:20:10 split.

Model	Configuration	Precision	Recall	Specificity	F1-Score
EfficientNet-B0	Augmentation + Gaussian	0.9845	0.9621	0.9848	0.9732
EfficientNet-B0	Augmentation Only	0.9580	0.8636	0.9621	0.9084
EfficientNet-B0	No Augmentation	0.9466	0.9394	0.9470	0.9430
MobileNetV2	Augmentation + Gaussian	0.9048	0.8636	0.9091	0.8837
MobileNetV2	Augmentation Only	0.8601	0.9318	0.8485	0.8945

smoother, the variance analysis shows only a small difference between the two configurations.

For MobileNetV2, validation performance was generally lower and more variable than that of EfficientNet-B0. Under the 70:20:10 configuration, MobileNetV2 achieved a mean validation accuracy of 87.80% with a standard deviation of 1.32% (variance = 0.00017). Meanwhile, the 80:10:10 configuration produced a lower mean accuracy of 85.96% with substantially higher variability, indicated by a standard deviation of 3.62% and variance of 0.00131. These findings suggest that MobileNetV2 is more sensitive to changes in data distribution and in data-splitting strategies. In addition, the fold-to-fold fluctuations observed in Fig. 6(c) and Fig. 6(d) indicate that MobileNetV2 converges less stably during the early training epochs. This behavior is particularly evident in the 80:10:10 configuration, where several folds show inconsistent validation trends before convergence. Overall, the experimental results demonstrate that both data splitting strategies enable effective model learning. However, EfficientNet-B0 consistently achieves higher validation accuracy and lower variability compared to MobileNetV2. The 70:20:10 configuration provides the most balanced and stable performance, particularly for EfficientNet-B0, indicating that a larger validation portion contributes to more reliable performance estimation during training.

B. Cross-Validation and Statistical Analysis

To further examine model consistency and reliability, five-fold cross-validation was applied to all experimental configurations. The evaluation focused on the average validation accuracy and fold variability across training epochs, as presented in Fig. 7(a)–(d). EfficientNet-B0 demonstrates consistently strong performance under both data partitioning strategies. In the 70:20:10 configuration, the model achieved a mean validation accuracy of 95.98% with a standard deviation of 0.99%, while in the 80:10:10 configuration, it achieved 95.80% with a standard deviation of 1.34%. The relatively small deviation values indicate that EfficientNet-B0 maintains stable performance across all folds, reflecting good reproducibility and reliable generalization capability. The variance analysis in Table 6 further supports these findings. EfficientNet-B0 with the 70:20:10 split and augmentation plus Gaussian preprocessing produced the

Table. 6 Variance Comparison of Accuracy Across Folds

Model	Configuration	Mean Accuracy	Std. Dev	Variance
EfficientNet-B0	70:20:10, Aug + Gaussian	0.9598	0.0099	0.00010
EfficientNet-B0	80:10:10, Aug + Gaussian	0.9580	0.0134	0.00018
EfficientNet-B0	70:20:10, Aug only	0.9311	0.0098	0.0098
EfficientNet-B0	70:20:10, No Aug	0.9568	0.0131	0.00017
MobileNetV2	70:20:10, Aug + Gaussian	0.8780	0.0132	0.00017
MobileNetV2	80:10:10, Aug + Gaussian	0.8596	0.0362	0.00131
MobileNetV2	70:20:10, Aug only	0.8659	0.0533	0.00284

Corresponding author: Yunidar, yunidar@usk.ac.id, Department of Electrical Engineering and Computer, Universitas Syiah Kuala, Banda Aceh, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.338

lowest variance value of 0.00010, whereas the 80:10:10 configuration resulted in a slightly higher variance of 0.00018. These results indicate that the 70:20:10 configuration provides more consistent validation performance across repeated experiments.

In contrast, MobileNetV2 shows greater variability across folds. The 70:20:10 configuration achieved a mean validation accuracy of 87.80% with a standard deviation of 1.32%, while the 80:10:10 configuration achieved 85.96% with a substantially larger standard deviation of 3.62%. The higher fluctuation indicates that MobileNetV2 is more sensitive to differences in the composition of training and validation data. This trend is also reflected in the variance values. MobileNetV2 with the 80:10:10 configuration produced a variance of 0.00131, while the augmentation-only configuration reached the highest variance value of 0.00284. Compared to EfficientNet-B0, these results demonstrate lower stability and less consistent fold performance for MobileNetV2. To determine whether the observed performance differences were statistically significant, the Friedman test was conducted across all configurations. As summarized in Table 7, EfficientNet-B0 with augmentation and Gaussian preprocessing achieved Friedman statistics of 49.9313 for the 70:20:10 configuration and 40.3117 for the 80:10:10 configuration, both with p-values < 0.001. Similarly, MobileNetV2 achieved statistically significant results with Friedman statistics of 16.9728 and 30.1010 for the 70:20:10 and 80:10:10 configurations, respectively. The statistical results confirm that the preprocessing strategy and data partitioning significantly influence model performance. EfficientNet-B0 consistently achieved higher statistical scores and lower variability, indicating superior robustness across folds. Meanwhile, MobileNetV2 maintained competitive classification capability but exhibited greater sensitivity to data distribution changes. Overall, the combination of cross-validation analysis and statistical testing demonstrates that EfficientNet-B0 with the 70:20:10 configuration provides the most stable and reliable performance among all evaluated configurations.

C. Comparative Performance Analysis

Table 4 summarizes the comparative performance of EfficientNet-B0 and MobileNetV2 under the 70:20:10 and 80:10:10 data split configurations using augmentation combined with Gaussian preprocessing. Among all evaluated configurations, EfficientNet-B0 with the 70:20:10 split achieved the best overall performance, obtaining a precision of 0.9845, recall of 0.9621, specificity of 0.9848, and F1-score of 0.9732. These values indicate that the model can correctly distinguish DS and non-DS facial images with high sensitivity and low misclassification rates. When the data split was changed to 80:10:10, EfficientNet-B0 showed a decrease in performance, with precision, recall, specificity, and F1-score declining to 0.9603, 0.9237, 0.9618, and 0.9416, respectively. The approximately 3.16% reduction in F1-score suggests that increasing the training proportion does not necessarily improve overall classification reliability. Instead, the smaller validation portion may reduce the model's ability to generalize consistently across folds.

Table. 7 Friedman Test Results for Model Performance Comparison Across Configurations

Model	Configuration	Statistic	p-value	Significance
EfficientNet-B0	70:20:10, Aug + Gaussian	49.9313	0.000000	Significant
EfficientNet-B0	80:10:10, Aug + Gaussian	40.3117	0.000000	Significant
EfficientNet-B0	70:20:10, Aug only	44.3471	0.000000	Significant
EfficientNet-B0	70:20:10, No Aug	4.5397	0.000005	Not Significant
MobileNetV2	70:20:10, Aug + Gaussian	16.9728	0.0132	Significant
MobileNetV2	80:10:10, Aug + Gaussian	30.1010	0.000005	Significant
MobileNetV2	70:20:10, Aug only	0.000144	0.000144	Significant

Compared to EfficientNet-B0, MobileNetV2 produced lower performance across both configurations. Under the 70:20:10 split, MobileNetV2 achieved a precision of 0.9048, a recall of 0.8636, a specificity of 0.9091, and an F1-score of 0.8837. Although the model maintained acceptable classification capability, all performance metrics remained lower than those obtained by EfficientNet-B0. The weakest performance was observed in MobileNetV2 using the 80:10:10 configuration. While the recall increased to 0.9695, the specificity decreased substantially to 0.6947, indicating that the model incorrectly classified a larger number of non-DS samples as DS. This imbalance caused the F1-score to decrease to 0.8523, demonstrating reduced classification stability and poorer class discrimination capability. The confusion matrix analysis in Fig. 8 further supports these findings. As shown in Fig. 8(a), EfficientNet-B0 with the 70:20:10 configuration produced only 2 false positives and 5 false

Corresponding author: Yunidar, yunidar@usk.ac.id, Department of Electrical Engineering and Computer, Universitas Syiah Kuala, Banda Aceh, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.338

negatives. In contrast, Fig. 8(b) shows that the 80:10:10 configuration resulted in 5 false positives and 10 false negatives, indicating a higher number of classification errors. The increase in false negatives is particularly significant because undetected DS cases may reduce the effectiveness of early screening systems. Although the 80:10:10 configuration achieved competitive accuracy, the confusion matrix results demonstrate that the 70:20:10 configuration provides a more balanced trade-off between sensitivity and specificity, resulting in more reliable classification performance. Overall, the comparative analysis confirms that EfficientNet-B0 consistently outperforms MobileNetV2 across all evaluation metrics and data split strategies. In addition, the 70:20:10 configuration provides the most balanced and stable performance, achieving higher F1-scores, lower misclassification rates, and greater consistency across validation folds.

D. Ablation Study on Preprocessing Techniques

This section evaluates the influence of preprocessing techniques, specifically data augmentation and Gaussian noise, on the classification performance of EfficientNet-B0 and MobileNetV2. To ensure a controlled comparison, all ablation experiments were conducted using the 70:20:10 data split configuration, which previously demonstrated the most stable performance. The experimental results are summarized in Table 5. For EfficientNet-B0, the combination of augmentation and Gaussian noise produced the best overall performance, achieving a precision of 0.9845, a recall of 0.9621, a specificity of 0.9848, and an F1-score of 0.9732. When Gaussian noise was removed, and only augmentation was applied, the F1-score decreased substantially to 0.9084, while recall also declined to 0.8636. This represents a 6.48% reduction in F1-score, indicating that Gaussian noise significantly improves feature robustness and model generalization.

The configuration without augmentation performed worse than the augmented configurations. EfficientNet-B0 without augmentation achieved a precision of 0.9466, a recall of 0.9394, a specificity of 0.9470, and an F1-score of 0.9430. Although the results remain relatively high, they are still inferior to the augmentation plus Gaussian configuration, suggesting that preprocessing techniques help reduce overfitting and improve classification consistency. For MobileNetV2, the influence of preprocessing was less consistent. Using augmentation with Gaussian noise, the model achieved precision of 0.9048, recall of 0.8636, specificity of 0.9091, and an F1-score of 0.8837. In contrast, the augmentation-only configuration produced a slightly higher F1-score of 0.8945 and a higher recall of 0.9318. However, this improvement was accompanied by a decrease in specificity to 0.8485, indicating an increase in false positive predictions. These results indicate that MobileNetV2 is more sensitive to preprocessing variations and less robust to Gaussian perturbations compared to EfficientNet-B0. Although the recall improved under augmentation-only preprocessing, the reduced specificity demonstrates weaker class discrimination capability and less balanced classification performance.

To further evaluate the consistency of these performance differences, the Friedman test was applied across all experimental configurations, as summarized in Table 7. EfficientNet-B0 with augmentation and Gaussian preprocessing achieved the highest Friedman statistic values, namely 49.9313 for the 70:20:10 configuration and 40.3117 for the 80:10:10 configuration, both with p-values < 0.001. These results indicate statistically significant and highly consistent performance improvements across cross-validation folds. Similarly, MobileNetV2 configurations also produced statistically significant results, including Friedman statistics of 16.9728 ($p = 0.0132$) for the 70:20:10 configuration and 30.1010 ($p < 0.001$) for the 80:10:10 configuration. However, the lower statistical values compared to EfficientNet-B0 indicate higher variability and reduced consistency across folds. The configuration without augmentation produced the weakest statistical result and did not demonstrate consistent performance improvement across folds. This finding confirms that preprocessing techniques, particularly augmentation combined with Gaussian noise, play an important role in stabilizing model learning and improving classification robustness, especially when working with limited datasets. Overall, the ablation study demonstrates that preprocessing strategies substantially influence model performance and stability. The combination of augmentation and Gaussian noise consistently provides the best and most reliable results, particularly for EfficientNet-B0, by improving feature robustness, reducing overfitting, and enhancing classification consistency across cross-validation folds.

IV. Discussion

A. Model Performance Analysis

This study evaluates the effect of two data partitioning strategies, namely 70:20:10 and 80:10:10, on the classification performance of EfficientNet-B0 for Down Syndrome facial image classification. The evaluation was conducted using five-fold cross-validation to ensure reliable performance estimation across different training and validation subsets. Based on the experimental results in Table 4, EfficientNet-B0 with the 70:20:10 configuration achieved precision of 0.9845, recall of 0.9621, specificity of 0.9848, and an F1-score of 0.9732. In comparison, the 80:10:10 configuration produced lower performance values, with a precision of 0.9603, recall of 0.9237, specificity of 0.9618, and F1-score of 0.9416. The approximately 3.16% decrease in F1-score indicates that increasing the proportion of training data does not necessarily improve classification reliability. The cross-validation analysis further demonstrates that the 70:20:10 configuration provides more stable performance across folds. As reported in Table 6, this configuration achieved a mean validation accuracy of 95.98% with a standard deviation of 0.99% and variance of 0.00010, whereas the 80:10:10 configuration achieved a mean validation accuracy of 95.80% with a higher standard deviation of

Corresponding author: Yunidar, yunidar@usk.ac.id, Department of Electrical Engineering and Computer, Universitas Syiah Kuala, Banda Aceh, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.338

1.34% and variance of 0.00018. These findings indicate that the larger validation portion in the 70:20:10 configuration contributes to more consistent model evaluation and reduced fold-to-fold variability. The confusion matrix analysis shown in Fig. 8 also supports these findings. The 70:20:10 configuration produced only 2 false positives and 5 false negatives, while the 80:10:10 configuration resulted in 5 false positives and 10 false negatives. The higher number of false negatives in the 80:10:10 configuration indicates that more DS cases were incorrectly classified as non-DS. From a clinical perspective, minimizing false negatives is particularly important because undetected DS cases may delay early diagnosis and intervention. Therefore, although the 80:10:10 configuration achieved competitive accuracy, the 70:20:10 configuration provides more balanced classification performance and better reliability for screening applications.

The Friedman test results presented in Table 7 further confirm that the observed performance differences are statistically significant ($p < 0.05$). EfficientNet-B0 with augmentation and Gaussian preprocessing achieved Friedman statistics of 49.9313 for the 70:20:10 configuration and 40.3117 for the 80:10:10 configuration, both with highly significant p-values (< 0.001). These results indicate that the differences in performance are not caused by random variation but are influenced by the selected data partitioning strategy and preprocessing approach. From a machine learning perspective, these findings reflect the bias-variance trade-off. The 80:10:10 configuration benefits from a larger training dataset, which may improve learning capability and reduce bias. However, the smaller validation set increases sensitivity to data distribution changes, resulting in higher variability across folds. In contrast, the 70:20:10 configuration provides a more balanced allocation between training and validation data, leading to improved stability and more reliable performance estimation. Overall, the discussion demonstrates that model evaluation should not focus solely on achieving the highest accuracy value. Stability, consistency across folds, and balanced error distribution are equally important, particularly in medical image classification tasks where reliable prediction performance is required for practical clinical applications.

B. Comparison with Previous Study

The findings of this study are consistent with previous research regarding the influence of data partitioning strategies on machine learning performance. Bichri et al. [17] reported that increasing the proportion of training data beyond 70% can improve classification performance; however, the improvement is generally limited and strongly influenced by dataset characteristics and evaluation methodology. Similarly, Rácz et al. [16] demonstrated that dataset quality and size have a greater impact on model performance than the train-test split ratio itself. Their study emphasized that allocating a larger portion of data for training does not always guarantee better generalization, particularly when the validation or testing subsets become too limited for reliable evaluation.

The results obtained in this study support these observations. Although the 80:10:10 configuration provided competitive performance, the improvement in validation accuracy compared to the 70:20:10 configuration was relatively small. EfficientNet-B0 achieved a mean validation accuracy of 95.80% under the 80:10:10 configuration, compared to 95.98% for the 70:20:10 configuration. Furthermore, the 80:10:10 configuration produced higher variability, with a standard deviation of 1.34% and a variance of 0.00018, whereas the 70:20:10 configuration achieved a lower standard deviation of 0.99% and a variance of 0.00010. These findings indicate that increasing the training proportion does not necessarily improve performance stability. This study also extends previous work by employing a more comprehensive evaluation framework that integrates five-fold cross-validation and Friedman statistical testing. Unlike conventional single-split evaluation approaches, repeated cross-validation provides more reliable performance estimates because the model is evaluated across multiple training and validation subsets. In addition, the Friedman test is more appropriate for cross-validation analysis because the observations across folds are not fully independent.

The statistical analysis confirms that the observed performance differences are significant ($p < 0.05$), particularly for EfficientNet-B0 with augmentation and Gaussian preprocessing, which achieved Friedman statistics of 49.9313 for the 70:20:10 configuration and 40.3117 for the 80:10:10 configuration. These results demonstrate that the differences between configurations are not caused by random variation but are associated with the selected data partitioning and preprocessing strategies. Compared with previous studies, the present work provides additional insight into the relationships among data splitting strategy, model stability, and classification reliability in medical image analysis. The results indicate that the 70:20:10 configuration offers more balanced and consistent performance, particularly in reducing fold variability and minimizing false negative predictions. This is particularly important in Down Syndrome screening applications, where stable and reliable classification performance is required to support early detection. Overall, this study confirms that model evaluation should consider not only average accuracy but also performance consistency, statistical significance, and error distribution. These findings contribute to a more comprehensive understanding of how data partitioning strategies influence deep learning performance in medical image classification tasks.

C. Research Limitations

Despite achieving promising classification performance, several limitations of this study should be acknowledged. First, although five-fold cross-validation was implemented to improve evaluation reliability, the dataset remains relatively limited in diversity because all facial images were collected from a single publicly available source. After the filtering and quality control stages, only 2,620 images were used for experimentation. For deep learning applications, this dataset size is still relatively small and

Corresponding author: Yunidar, yunidar@usk.ac.id, Department of Electrical Engineering and Computer, Universitas Syiah Kuala, Banda Aceh, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.338

may limit the model's ability to learn more complex facial feature variations. As a result, the generalization capability of the proposed model to broader real-world populations may remain limited. Another limitation is related to the absence of detailed subject-level metadata. The dataset does not provide comprehensive information regarding ethnicity, geographic background, acquisition conditions, or subject identity. Consequently, subject-wise data splitting could not be performed, creating a potential risk of identity overlap between training and testing subsets. Although duplicate filtering was conducted during preprocessing, complete elimination of identity-level similarity could not be guaranteed. This limitation may lead to optimistic performance estimation and should therefore be interpreted carefully.

The study also evaluates only two data partitioning strategies, namely 70:20:10 and 80:10:10. While the results demonstrate that the 70:20:10 configuration provides more stable performance, other partitioning methods, such as stratified cross-validation, nested cross-validation, or subject-independent validation, were not investigated. Exploring additional validation approaches may provide a more comprehensive understanding of optimal data partitioning strategies for medical image classification. In terms of model architecture, this research primarily focuses on EfficientNet-B0 and MobileNetV2. Although EfficientNet-B0 achieved the best overall performance, with an F1-score of 0.9732 and variance of 0.00010 under the 70:20:10 configuration, comparisons with more recent architectures or transformer-based models were not included. Additional experiments using larger and more advanced architectures may provide further insight into performance scalability and robustness. The preprocessing strategy used in this study also remains relatively limited. While augmentation and Gaussian noise successfully improved classification performance, more sophisticated augmentation approaches, such as generative adversarial networks (GANs) or synthetic facial image generation, were not explored. These methods may help increase data diversity and improve robustness, particularly for limited medical datasets.

In addition to technical limitations, ethical considerations are important when applying facial image analysis for Down Syndrome screening. Although the dataset was obtained from publicly accessible sources, facial images still contain sensitive biometric information that may raise concerns regarding privacy, consent, and responsible data usage. Furthermore, the lack of demographic diversity in the dataset may introduce bias and reduce model fairness across different populations. From a practical perspective, the proposed system should not be considered a definitive diagnostic tool. The increase in false negatives observed in certain configurations, particularly under the 80:10:10 split, indicates that some DS cases may remain undetected. Therefore, this approach is more appropriate as a supportive screening system to assist healthcare professionals rather than as a replacement for clinical diagnosis. Future work should focus on collecting larger multi-source datasets, implementing subject-independent evaluation protocols, exploring more advanced augmentation techniques, and improving fairness across demographic groups. These improvements are necessary to enhance model generalization, reliability, and readiness for real-world clinical implementation.

D. Implications and Practical Significance

The results of this study demonstrate that the data partitioning strategy plays an important role in determining both classification performance and evaluation stability in deep learning-based medical image analysis. Although increasing the amount of training data is generally expected to improve feature learning capability, the experimental findings show that higher training proportions do not always produce more reliable model performance. In this study, EfficientNet-B0 using the 80:10:10 configuration achieved competitive performance with a mean validation accuracy of 95.80%. However, this configuration also produced higher variability, indicated by a standard deviation of 1.34% and a variance of 0.00018. In comparison, the 70:20:10 configuration achieved a slightly higher mean validation accuracy of 95.98%, while maintaining lower variability with a standard deviation of 0.99% and variance of 0.00010. These findings indicate that a more balanced allocation between training and validation data contributes to more stable and consistent model evaluation. The practical implications of these results are particularly important in medical screening applications. Based on the confusion matrix analysis, the 70:20:10 configuration generated only 2 false positives and 5 false negatives, whereas the 80:10:10 configuration produced 5 false positives and 10 false negatives. The increase in false negatives is clinically significant because undetected Down Syndrome cases may delay early diagnosis and intervention. Therefore, despite the competitive accuracy achieved by the 80:10:10 configuration, the 70:20:10 configuration provides more balanced and clinically reliable classification performance.

The findings also indicate that evaluation stability should be considered together with overall accuracy. A larger training subset may improve learning capability, but a smaller validation subset can increase sensitivity to fold variability and reduce the reliability of performance estimation. This observation reflects the bias-variance trade-off in machine learning, where reducing bias through larger training data may simultaneously increase variance across validation folds. Another important contribution of this study is the implementation of a more rigorous evaluation framework through the integration of five-fold cross-validation and Friedman statistical testing. The Friedman test is particularly appropriate for repeated cross-validation experiments because it does not assume independence between observations across folds. In this study, the Friedman analysis produced statistically significant results ($p < 0.05$) for most experimental configurations, confirming that the observed differences in performance were not caused by random variation. From a methodological perspective, the combination of cross-validation, variance analysis, confusion matrix evaluation,

Corresponding author: Yunidar, yunidar@usk.ac.id, Department of Electrical Engineering and Computer, Universitas Syiah Kuala, Banda Aceh, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.338

and non-parametric statistical testing provides a more comprehensive and reliable framework for evaluating deep learning models in medical image classification. This evaluation approach improves the reliability, reproducibility, and interpretability of the experimental findings, making the results more suitable for real-world clinical applications. Overall, this study highlights that the evaluation of medical AI systems should not focus solely on achieving the highest accuracy value. Performance consistency, balanced error distribution, statistical significance, and evaluation stability are equally important to ensure that AI-based screening systems can provide reliable and clinically meaningful support in practical healthcare environments.

V. Conclusion

This study evaluated the performance of an EfficientNet-B0-based model for Down Syndrome (DS) facial image classification, with a particular emphasis on comparing two data splitting strategies: 70:20:10 and 80:10:10. Experimental results demonstrate that both configurations achieved strong classification performance, with average accuracies exceeding 87% across five-fold cross-validation. The 80:10:10 data split consistently yielded higher peak performance, achieving a maximum accuracy of 92.36% and a higher average F1-score than the 70:20:10 configuration, while also exhibiting lower fold-to-fold performance variability. Statistical analysis using the Friedman test confirmed that performance differences among configurations are statistically significant ($p < 0.05$), which is more appropriate for cross-validation-based evaluation.

Declarations

Author Contributions

Yunidar conceived the framework, designed the methodology, supervised, and reviewed the manuscript. Melinda and Dzaky Dhiya Ul-Haq managed data collection, curation, preprocessing, and experimental setup. Rosmawinda developed, trained, validated, and evaluated models using EfficientNet-B0 and MobileNetV2. Nurlida Basir conducted statistical analyses, including cross-validation, variance analysis, and the Friedman test. All authors contributed to and approved the final manuscript.

Funding

This research received no external funding.

Data Availability

The facial image dataset used in this study was obtained from publicly available sources through the Roboflow platform. Processed data, trained model outputs, and additional experimental results are available from the corresponding author upon reasonable request.

Use of Artificial Intelligence

The authors declare that no artificial intelligence (AI) tools were used in the design, analysis, interpretation, writing, or preparation of this manuscript.

Acknowledgments

The authors would like to express their sincere gratitude to Universitas Syiah Kuala for providing academic support and research facilities. The authors also thank the providers of the publicly available facial image dataset through the Roboflow platform, which made this research possible. Appreciation is extended to all individuals and institutions who directly or indirectly contributed to the completion of this study.

Ethical Approval

This study utilized publicly available facial image datasets obtained from the Roboflow platform and other publicly accessible sources. No direct interaction with human participants, clinical intervention, or collection of personally identifiable information was conducted by the authors. Although formal ethical approval was not required under institutional guidelines for the use of publicly available secondary data, the study was conducted in accordance with ethical principles for research involving human-related data. Special consideration was given to privacy, responsible data usage, and the protection of sensitive biometric information contained in facial images. The authors acknowledge the potential ethical concerns associated with facial image analysis and emphasize that the proposed approach is intended solely for research and supportive screening purposes, not as a substitute for clinical diagnosis.

Consent for Publication Participants.

All authors have read and approved the final manuscript and consent to its publication.

References

[1] L. Marques Silva, M. Lucia Pereira Da, E. Tenorio, and V. Mattos De Oliveira, “Living and Learning with Down Syndrome,” *Int. J. Sci. Res. IJSR*, vol. 10, no. 7, pp. 1117–1124, Jul. 2021, doi: 10.21275/SR21427004518.

[2] S. Khurana, A. R. Khalifa, N. N. Rezallah, S. Lozanoff, and A. Z. Abdelkarim, “Craniofacial and Airway Morphology in Down Syndrome: A Cone Beam Computed Tomography Case Series Evaluation,” *J. Clin. Med.*, vol. 13, no. 13, p. 3908, Jul. 2024, doi: 10.3390/jcm13133908.

[3] J. Shetty, A. Shetty, S. C. Mundkur, T. K. Dinesh, and P. Pundir, “Economic burden on caregivers or parents with Down syndrome children—a systematic review protocol,” *Syst. Rev.*, vol. 12, no. 1, p. 3, Jan. 2023, doi: 10.1186/s13643-022-02165-2.

[4] A. Utari, F. K. Cayami, T. A. Rahardjo, S. E. Sabatini, V. Ulvyana, and T. I. Winarni, “Critical issue in the identification of Down syndrome and its problems in

Corresponding author: Yunidar, yunidar@usk.ac.id, Department of Electrical Engineering and Computer, Universitas Syiah Kuala, Banda Aceh, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.338

Central Java, Indonesia: The fact of needing health care and better management," *Intractable Rare Dis. Res.*, vol. 13, no. 2, pp. 121–125, May 2024, doi: 10.5582/irdr.2023.01103.
[5] E. Fucà, F. Costanzo, L. Celestini, A. Mandarino, and S. Vicari, "Characterization of Sleep Disturbances in Children and Adolescents with Down Syndrome and Their Relation with Cognitive and Behavioral Features," *Int. J. Environ. Res. Public. Health*, vol. 18, no. 9, p. 5001, May 2021, doi: 10.3390/ijerph18095001.
[6] O. A. Ijezie, J. Healy, P. Davies, E. Balaguer-Ballester, and V. Heaslip, "Quality of life in adults with Down syndrome: A mixed methods systematic review," *PLOS ONE*, vol. 18, no. 5, p. e0280014, May 2023, doi: 10.1371/journal.pone.0280014.
[7] U. Bellugi, H. Sabo, and J. Vaid, "Spatial Deficits in Children with Williams Syndrome," in *Spatial Cognition*, 1st ed., New York: Psychology Press, 2022, pp. 273–298. doi: 10.4324/9781315785462-16.
[8] Y. Yunidar, M. Melinda, and M. Irhamsyah, "Object Segmentation in Stunted Face Images using Deeplabv3+ with Resnet-50," *J. Nas. Tek. Elektro*, pp. 137–143, Nov. 2024, doi: 10.25077/jnte.v13n3.1253.2024.
[9] Y. Yunidar, Y. Yusni, N. Nasaruddin, and F. Arnia, "CNN Performance Improvement for Classifying Stunted Facial Images Using Early Stopping Approach," *J. RESTI Rekayasa Sist. Dan Teknol. Inf.*, vol. 9, no. 1, pp. 62–68, Jan. 2025, doi: 10.29207/resti.v9i1.6068.
[10] Y. Yunidar, R. Roslidar, M. Oktiana, Y. Yusni, N. Nasaruddin, and F. Arnia, "Classification of stunted and normal children using novel facial image database and convolutional neural network," *Radioelectron. Comput. Syst.*, vol. 2024, no. 1, pp. 76–86, Feb. 2024, doi: 10.32620/reks.2024.1.07.
[11] Y. Yunidar *et al.*, "CNN-Based Facial Image Analysis for Pediatric Down Syndrome Classification," *J. Electron. Electromed. Eng. Med. Inform.*, vol. 8, no. 2, pp. 696–711, Apr. 2026, doi: 10.35882/jeeemi.v8i2.1523.
[12] M. D. R. Kasha, Y. Yunidar, M. Melinda, N. Basir, and S. Rusdiana, "Robust Facial Classification of Down Syndrome using Lightweight CNNs," *J. INFOTEL*, vol. 18, no. 1, 2026, doi: 10.20895/infotel.v18i1.1525.
[13] Y. Gulzar, "Fruit Image Classification Model Based on MobileNetV2 with Deep Transfer Learning Technique," *Sustainability*, vol. 15, no. 3, p. 1906, Jan. 2023, doi: 10.3390/su15031906.
[14] M. Akay *et al.*, "Deep Learning Classification of Systemic Sclerosis Skin Using the MobileNetV2 Model," *IEEE Open J. Eng. Med. Biol.*, vol. 2, pp. 104–110, 2021, doi: 10.1109/OJEMB.2021.3066097.
[15] I. Ramadhan *et al.*, "Mobile Application Development for Facial Classification of Autistic Children Based on MobileNet-V3," *J. INFOTEL*, vol. 17, no. 3, pp. 612–626, Sep. 2025, doi: 10.20895/infotel.v17i3.1363.
[16] A. Rácz, D. Bajusz, and K. Héberger, "Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification," *Molecules*, vol. 26, no. 4, p. 1111, Feb. 2021, doi: 10.3390/molecules26041111.
[17] H. Bichri, A. Chergui, and M. Hain, "Investigating the Impact of Train / Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 2, 2024, doi: 10.14569/IJACSA.2024.0150235.
[18] R. T. Turksoy and B. Turkmen, "The Effects of Data Split Strategies on the Offline Experiments for CTR Prediction," Jun. 26, 2024, *arXiv*: arXiv:2406.18320. doi: 10.48550/arXiv.2406.18320.
[19] P. Eko Niti Taruno, G. Satya Nugraha, R. Dwiyansaputra, and F. Bimantoro, "Monkeypox Classification based on Skin Images using CNN: EfficientNet-B0," *E3S Web Conf.*, vol. 465, p. 02031, 2023, doi: 10.1051/e3sconf/202346502031.
[20] N. Tanwar and A. V. Turukmane, "Modified MobileNetV2 transfer learning model to detect road potholes," *PeerJ Comput. Sci.*, vol. 11, p. e2519, Jan. 2025, doi: 10.7717/peerj-cs.2519.
[21] S. A. Zaidi, V. Chouvatut, C. Phongnarisorn, and D. Praserttitipong, "Deep learning based detection of endometriosis lesions in laparoscopic images with 5-fold cross-validation," *Intell.-Based Med.*, vol. 11, p. 100230, 2025, doi: 10.1016/j.ibmed.2025.100230.
[22] J. Liu and Y. Xu, "T-Friedman Test: A New Statistical Test for Multiple Comparison with an Adjustable Conservativeness Measure," *Int. J. Comput. Intell. Syst.*, vol. 15, no. 1, p. 29, Dec. 2022, doi: 10.1007/s44196-022-00083-8.
[23] "Down syndrome - v 2 2025-05-25 5:17pm > Images," Roboflow. Accessed: Sep. 23, 2025. [Online]. Available: https://universe.roboflow.com/ahmad-shahab/down-syndrome-q3cnc
[24] D. Stojnev and A. Stojnev Ilić, "Preprocessing Image Data for Deep Learning," in *Proceedings of the International Scientific Conference - Sinteza 2020*, Beograd, Serbia: Singidunum University, 2020, pp. 312–317. doi: 10.15308/Sinteza-2020-312-317.
[25] T. Sarsembayeva, M. Mansurova, A. Abdildayeva, and S. Serebryakov, "Enhancing U-Net Segmentation Accuracy Through Comprehensive Data Preprocessing," *J. Imaging*, vol. 11, no. 2, p. 50, Feb. 2025, doi: 10.3390/jimaging11020050.
[26] R. S. Thakur, S. Chatterjee, R. N. Yadav, and L. Gupta, "Image De-Noising With Machine Learning: A Review," *IEEE Access*, vol. 9, pp. 93338–93363, 2021, doi: 10.1109/ACCESS.2021.3092425.
[27] N. Ikizler and G. Ekim, "Investigating the effects of Gaussian noise on epileptic seizure detection: The role of spectral flatness, bandwidth, and entropy," *Eng. Sci. Technol. Int. J.*, vol. 64, p. 102005, Apr. 2025, doi: 10.1016/j.jestch.2025.102005.
[28] G. K. M and A. D. Goswami, "Automatic Classification of the Severity of Knee Osteoarthritis Using Enhanced Image Sharpening and CNN," *Appl. Sci.*,

Corresponding author: Yunidar, yunidar@usk.ac.id, Department of Electrical Engineering and Computer, Universitas Syiah Kuala, Banda Aceh, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.338

vol. 13, no. 3, p. 1658, Jan. 2023, doi: 10.3390/app13031658.
[29] T. Cevik, N. Cevik, M. S. Ag, O. Guler, and S. Saglam, "A unified deep learning framework for image denoising and sharpening: from sequential to end-to-end models," *Neural Comput. Appl.*, vol. 38, no. 8, p. 269, Apr. 2026, doi: 10.1007/s00521-026-12015-0.
[30] A. Rofena *et al.*, "A deep learning approach for virtual contrast enhancement in Contrast Enhanced Spectral Mammography," *Comput. Med. Imaging Graph.*, vol. 116, p. 102398, Sep. 2024, doi: 10.1016/j.compmedimag.2024.102398.
[31] R. Archana and P. S. E. Jeevaraj, "Deep learning models for digital image processing: a review," *Artif. Intell. Rev.*, vol. 57, no. 1, p. 11, Jan. 2024, doi: 10.1007/s10462-023-10631-z.
[32] S. Yang, W. Xiao, M. Zhang, S. Guo, J. Zhao, and F. Shen, "Image Data Augmentation for Deep Learning: A Survey," Nov. 05, 2023, *arXiv*: arXiv:2204.08610. doi: 10.48550/arXiv.2204.08610.
[33] Q. H. Nguyen *et al.*, "Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil," *Math. Probl. Eng.*, vol. 2021, pp. 1–15, Feb. 2021, doi: 10.1155/2021/4832864.
[34] K. Kansal, T. B. Chandra, and A. Singh, "ResNet-50 vs. EfficientNet-B0: Multi-Centric Classification of Various Lung Abnormalities Using Deep Learning," *Procedia Comput. Sci.*, vol. 235, pp. 70–80, 2024, doi: 10.1016/j.procs.2024.04.007.
[35] T. Öznacar and N. Varol Kayapunar, "Advanced skin cancer prediction with medical image data using MobileNetV2 deep learning and optimized techniques," *Sci. Rep.*, vol. 15, no. 1, p. 28962, Aug. 2025, doi: 10.1038/s41598-025-14963-4.
[36] A. S. Mohammad, T. G. Jarullah, M. T. S. Al-Kaltakchi, J. Alshehabi Al-Ani, and S. Dey, "IoT-MFaceNet: Internet-of-Things-Based Face Recognition Using MobileNetV2 and FaceNet Deep-Learning Implementations on a Raspberry Pi-400," *J. Low Power Electron. Appl.*, vol. 14, no. 3, p. 46, Sep. 2024, doi: 10.3390/jlpea14030046.
[37] M. Reyad, A. M. Sarhan, and M. Arafa, "A modified Adam algorithm for deep neural network optimization," *Neural Comput. Appl.*, vol. 35, no. 23, pp. 17095–17112, Aug. 2023, doi: 10.1007/s00521-023-08568-z.
[38] C. Li, K. Liu, and S. Liu, "A Survey of Loss Functions in Deep Learning," *Mathematics*, vol. 13, no. 15, p. 2417, Jul. 2025, doi: 10.3390/math13152417.
[39] S. Sathyanarayanan, "Confusion Matrix-Based Performance Evaluation Metrics," *Afr. J. Biomed. Res.*, pp. 4023–4031, Nov. 2024, doi: 10.53555/AJBR.v27i4S.4345.
[40] Department of Computer Science and Informatics, University of Energy and Natural Resources, Sunyani, Ghana, I. K. Nti, O. Nyarko-Boateng, and J. Aning, "Performance of Machine Learning Algorithms with Different K Values in K-fold CrossValidation," *Int. J. Inf. Technol. Comput. Sci.*, vol. 13, no. 6, pp. 61–71, Dec. 2021, doi: 10.5815/ijitcs.2021.06.05.

Author Biography

Dzaky Dhiya Ul-Haq is an undergraduate student in the Department of Electrical and Computer Engineering, Faculty of Engineering, Syiah Kuala University, Indonesia. He began his studies in 2022 and is currently completing his final project research entitled "Analysis of Data Separation Strategy for Down Syndrome Facial Image Classification Using EfficientNet-B0." His research interests include computer vision, deep learning, and medical image analysis, with a focus on non-invasive AI-based screening systems. His work involved dataset curation, image pre-processing, model fine-tuning, and performance evaluation using cross-validation and statistical analysis. He can be contacted via email: dzaky.d@mhs.usk.ac.id

Yunidar has been a lecturer at the Faculty of Engineering, Department of Electrical and Computer Engineering, Syiah Kuala University (USK) since March 2000. She completed her bachelor's degree in Physics at Syiah Kuala University, Aceh, Indonesia, in 1997. In 2000, she obtained a Master of Engineering (M.T.) degree in Optoelectronics and Laser Application from the University of Indonesia, Jakarta, Indonesia. She completed her doctoral studies in the Doctoral Program of Engineering Science at the Graduate School of Syiah Kuala University on February 6, 2025. She is a member of the IEEE and has also been actively involved as a member of the Indonesian Electrical Engineering Higher Education Forum (FORTEI) for Region 1, Sumatra, since 2023. Her research focuses on biomedical applications. She can be contacted via email: yunidar@usk.ac.id

Melinda was born in Bireuen, Aceh, on June 10, 1979. She received a B. Eng degree from the Department of Electrical and Computer Engineering, Faculty of Engineering, Universitas Syiah Kuala, Banda Aceh, in 2002. She completed her master's degree at the Faculty of Electrical Department, University of Southampton, United Kingdom, with a concentration in field study of Radio Frequency Communication Systems in 2009. She completed her Doctoral degree at the Department of Electrical Engineering, Engineering Faculty of Universitas Indonesia in February 2018. She has been with the Department of Electrical Engineering, Faculty of Engineering, Universitas Syiah Kuala since 2002. She is also a member of IEEE. Her research interests include

Corresponding author: Yunidar, yunidar@usk.ac.id, Department of Electrical Engineering and Computer, Universitas Syiah Kuala, Banda Aceh, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.338

multimedia signal processing and fluctuation processing. She can be contacted at email: melinda@usk.ac.id.

Dr. Nurlida Basir is an Associate Professor at the Faculty of Science and Technology, Universiti Sains Islam Malaysia (USIM). She began her academic career at USIM in 2002 and has since been actively involved in teaching, research, and academic leadership. She holds a Diploma, a Bachelor's, and a Master in Computer Science from Universiti Teknologi Malaysia (UTM), and earned her Ph.D. in Computer Science from the University of Southampton, United Kingdom. Her research interests span across software engineering, cybersecurity, malware detection, signal processing, and artificial intelligence. Her research has been extensively published in prominent academic journals and conference proceedings. Alongside her research, Dr. Basir is a dedicated educator, mentoring both undergraduate and postgraduate students in computer science. Dr. Nurlida is a member of the Institute of Electrical and Electronics Engineers (IEEE), reflecting her active participation in the global academic and research community. She can be contacted at nurlida@usim.edu.my

Rosmawinda was born in Pekanbaru on May 17, 2001. She is currently a Master's student in Electrical Engineering at Syiah Kuala University, with interests in multimedia technology and biomedicine. During her undergraduate studies, she focused on multimedia technology, particularly facial recognition research, which served as the foundation for her master's-level research. As a student in the class of 2025, she actively participates in lectures. She continues to deepen her understanding of image processing, deep learning, and the application of intelligent technology in electronics. Her commitment is reflected in her ongoing efforts to combine theory with practice, resulting in relevant and innovative research. Dedicated to scientific development and applied skills, she is determined to make a significant contribution to technological advancement, particularly in the development of biometric systems and artificial intelligence. She can be reached via email: rosmawinda@mhs.usk.ac.id.

Corresponding author: Yunidar, yunidar@usk.ac.id, Department of Electrical Engineering and Computer, Universitas Syiah Kuala, Banda Aceh, Indonesia.
Digital Object Identifier (DOI): https://doi.org/10.35882/ijeeemi.v8i3.338