JURNAL INFOTEL
VOL. 18, NO. 1, FEBRUARY 2026, PP. 18–41. DOI:10.20895/INFOTEL.V18I1.1525

RESEARCH ARTICLE

Robust Facial Classification of Down Syndrome using Lightweight CNNs

Muhammad Dika Rafi Kasha[1], Yunidar Yunidar[2*], Melinda Melinda[3], Nurlida Basir [4], and Siti Rusdiana[5]

[1,2,3]Department of Electrical Engineering and Computer, Universitas Syiah Kuala, Indonesia

[4]Faculty of Science and Technology, Universiti Sains Islam Malaysia (USIM), Malaysia

[5]Department of Mathematics, Faculty of Mathematics Natural Sciences, Universitas Syiah Kuala, Indonesia

*Corresponding email: yunidar@usk.ac.id

Received: January 12, 2026; Revised: March 31, 2026; Accepted: April 06, 2026.

Abstract: Down syndrome (DS) is a genetic disorder caused by trisomy 21 and is associated with distinctive facial characteristics that can support early screening, particularly in resource-limited settings. This study aims to evaluate and compare the performance of convolutional neural network (CNN) architectures, EfficientNet-B1 and MobileNetV3-Large, for facial image-based DS classification, while enhancing model interpretability using Grad-CAM visualization. A dataset of 3,030 facial images derived from previously published studies was utilized, and after data cleaning and quality control, 2,620 images were used for model development. Image enhancement was applied only to the training data to avoid data leakage. The dataset was split into training, validation, and test sets with a 70:20:10 ratio, and both models were fine-tuned using ImageNet pre-trained weights. Model performance was evaluated using accuracy, precision, recall, and F1-score, while robustness was assessed through five-fold cross-validation. Performance differences were analyzed using one-way ANOVA. Experimental results indicate that EfficientNet-B1 achieved a higher average accuracy of 91.14%, compared to 88.56% for MobileNetV3-Large, with lower variability across validation folds. ANOVA analysis confirmed a statistically significant difference between the models ($p < 0.05$). Furthermore, Grad-CAM visualizations revealed that both models focused on clinically relevant facial regions, with EfficientNet-B1 demonstrating more consistent and interpretable attention patterns. These findings suggest that EfficientNet-B1 offers a robust and interpretable approach.

Keywords: ANOVA,classification, CNN, down-syndrome, models.

 ISSN: 2085-3688; e-ISSN: 2460-0997

1 Introduction

Down Syndrome (DS) is a genetic condition caused by an extra copy of chromosome 21. It is a leading cause of learning and developmental challenges in children. Studies show that early detection of DS is important for timely intervention, which can improve development and quality of life [1, 2]. In many areas, including Indonesia, where healthcare access can be difficult, early detection is often slowed by limited testing resources, a shortage of trained professionals, and differences in support services. These issues can delay help and affect a child's development [3–5]. Despite the importance of early identification, traditional DS screening approaches are constrained by the need for specialized tests, trained clinicians, and laboratory facilities. These limitations make consistent early assessment difficult, particularly in low-resource settings. This challenge underscores the need for more accessible, objective, and scalable screening methods that do not rely heavily on specialized infrastructure.

In recent years, deep learning has emerged as a powerful tool in medical and biometric image analysis. Convolutional neural networks (CNNs) are particularly effective because they can extract complex visual patterns directly from images, enabling them to distinguish individuals with DS from those without high reliability [6–9]. Earlier studies demonstrated that CNN-based classifiers can identify DS with strong performance [6], while other studies validated the ability of CNN-based models to classify DS and related syndromes [7]. More recent studies have combined modern architectures such as MobileNetV3 and ResNet to achieve accuracy above 99% across diverse datasets, highlighting the rapid advancement of DS-focused deep learning methods [10].

Deep learning has also been applied to prenatal imaging. Previous studies developed a CNN-based segmentation model for identifying nuchal translucency in ultrasound images, achieving a promising accuracy of early risk assessment of DS [11]. Although ultrasound and facial image approaches differ, both areas underscore the potential of deep learning for identifying DS-related visual characteristics. In addition to prenatal and facial imaging studies, several researchers have reported that CNN-based facial image analysis can also be effectively applied to other medical and developmental conditions. Facial morphology has been shown to contain discriminative visual features that can be learned reliably by CNN models for classifying conditions such as stunting and autism, using both conventional and lightweight architectures.

These studies indicate that deep learning-based facial analysis is not limited to a single disorder but has broader applicability for medical screening and early detection across different populations and datasets [12–15]. Furthermore, the use of an efficient CNN architecture supports practical implementation in real-world and resource-constrained environments, reinforcing the feasibility of deploying deep learning models for health-related facial images [12, 15, 16].

Among current architectures, MobileNetV3 and EfficientNet-B1 stand out for their efficiency and balanced performance. EfficientNet-B1 has been shown to perform well in facial recognition tasks due to its optimized scaling strategy [17], while MobileNetV3 is widely used for lightweight deployment scenarios because of its low computational requirements. These characteristics make both architectures suitable candidates for DS classification, especially in environments with limited computing resources. A common challenge in deep learning is the lack of interpretability, which can reduce clinician trust. To address this, explainable AI techniques such as Gradient-weighted Class Activation Mapping (Grad-

CAM) generate visual heatmaps highlighting which image regions influence the model's prediction. Such interpretability tools have been successfully applied in facial recognition and emotion analysis tasks, making them relevant for DS-related classification as well [18].

Despite the promising results reported in previous studies, several limitations remain that restrict the practical adoption of deep learning-based DS screening systems. Most existing approaches rely on complex or computationally intensive architectures, which limit their deployment in low-resource environments such as rural healthcare facilities or mobile-based screening platforms. Moreover, comparative analyses focusing specifically on lightweight CNN architectures for DS facial classification remain limited, making it difficult to identify models that provide an optimal balance between accuracy, stability, and computational efficiency.

Another critical limitation is the lack of interpretability analysis in many prior studies. While high classification accuracy is often reported, insufficient attention is given to explaining how models reach their decisions. In medical applications, this lack of transparency can reduce clinician trust and hinder real-world adoption. Although Grad-CAM has demonstrated effectiveness in visualizing discriminative regions in facial analysis tasks, its systematic integration into DS facial classification remains underexplored.

In addition to the aforementioned limitations, a significant concern in contemporary research is the insufficient examination of model stability and reliability. Many previous studies rely on a single train–test split, which may lead to overly optimistic performance estimates and fail to capture the inherent variability of medical image data. Moreover, limited efforts have been made to evaluate whether model performance remains consistent across different data partitions, thereby reducing the reliability and reproducibility of the reported results.

In the context of medical image classification, robustness and consistency are as critical as accuracy, since unstable predictions may lead to unreliable or unsafe clinical decisions. Therefore, a more rigorous evaluation strategy, incorporating cross-validation and statistical analysis, is necessary to ensure that model performance is both reliable and generalizable.

To address the identified research gaps, this study makes the following contributions:

1. **Structured Comparison of Lightweight CNN Architectures.** This study provides a systematic and fair comparison between two modern lightweight convolutional neural network architectures, MobileNetV3-Large and EfficientNet-B1, specifically for Down Syndrome facial classification. Unlike prior works that evaluate models independently or focus on complex architectures, this study offers a structured analysis of the trade-off between classification performance, stability, and computational efficiency. This comparison advances current knowledge by providing clearer guidance on model selection for DS classification in resource-constrained environments.
2. **Robustness and Statistical Reliability Evaluation.** To overcome the limitations of single train–test split evaluation, this study employs a more rigorous validation framework using 5-fold cross-validation. In addition, one-way Analysis of Variance (ANOVA) is applied to statistically assess performance consistency across folds. This contribution advances current practice by introducing a reliability-oriented evaluation approach, ensuring that model performance is not only accurate but also stable, reproducible, and less sensitive to data partitioning.
3. **Integration of Explainable AI for Model Interpretability.** This study integrates Grad-CAM as an explainability technique to visualize the important facial regions in-

fluencing model predictions. Unlike many previous DS classification studies that focus solely on accuracy, this work emphasizes interpretability to enhance transparency and support clinical trust. This contribution improves practical applicability by providing visual insights into the model's decision-making process, which is essential for real-world medical deployment.

In this study, we present a comparative analysis focusing on how two modern lightweight CNN architectures MobileNetV3 and EfficientNet-B1, perform in classifying facial images of individuals with and without DS. The primary objective is to systematically compare their classification effectiveness using accuracy, precision, recall, and F1-score, as well as their interpretability through a Grad-CAM-based analysis of prioritized facial regions. The models are trained on a dataset of 6,332 facial images obtained from the online platform, and split into 70% training, 20% validation, and 10% test sets to ensure robust performance assessment. To further enhance robustness and reduce bias, we use 5 fold cross validation. One-way Analysis of Variance (ANOVA) is applied to examine the statistical consistency of model performance across folds [19, 20].

This research provides a comprehensive comparison of MobileNetV3-Large and EfficientNet-B1 for facial-based DS classification, integrates Grad-CAM visual explanations to enhance model transparency and reliability, and proposes an accessible AI based screening framework that can support DS identification in settings with limited clinical expertise or diagnostic resources. Although the results demonstrate strong potential, limitations remain regarding dataset size and demographic diversity, which may affect generalizability. Future research should focus on collecting multi-center and multi-ethnic datasets to improve generalizability. Additional deep learning architectures and ensemble approaches can be explored. Integration of the model into a mobile-based screening application and clinical validation with medical experts are also important directions for further development.

2 Research Method

This study aims to compare two CNN architectures, MobileNetV3-Large and EfficientNet-B1, for classifying facial images of individuals with and without DS. The research workflow consists of five main stages: (1) data collection, (2) data preprocessing and augmentation, (3) model training and validation, (4) model evaluation and interpretability analysis, and (5) performance comparison between architectures. A total of 6,332 facial images from the public platform are used, with 70% for training, 20% for validation, and 10% for testing. Figure 1 summarizes the study's workflow, illustrating the sequential process from dataset preparation to performance comparison.

The overall workflow of this study is illustrated in Figure 1. The research process consists of several sequential stages designed to ensure systematic and reproducible experimentation. The workflow begins with the data collection stage, where facial images of individuals with and without DS are gathered and curated. This is followed by the preprocessing stage, which includes image enhancement, resizing, and partitioning the dataset into training, validation, and test subsets.

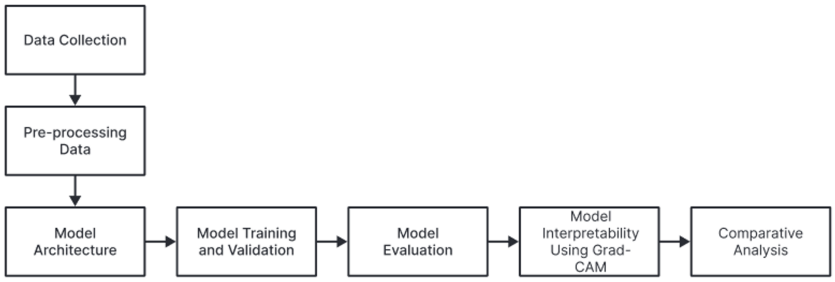

Figure 1: Research workflow.

2.1 Data Collection

The facial image dataset used in this study was compiled from publicly available, peer-reviewed sources and included labeled facial images of children with and without Down Syndrome (DS). This dataset has been used in several previous studies on face-based DS identification [9]. Specifically, the dataset referenced in comes from the Roboflow repository [21] developed for Down Syndrome detection, which contains facial images of children aged approximately 0–15 years, including individuals with Down syndrome and healthy controls.

All images were collected in accordance with ethical guidelines and anonymized to protect subject privacy. The dataset consists predominantly of frontal facial images captured under various lighting conditions and backgrounds, with a balanced distribution between DS and non-DS categories.

The facial image dataset used in this study was compiled from publicly available and peer-reviewed sources that provide labeled facial images of individuals with and without DS. The dataset was originally developed for research purposes and has been utilized in several previous studies on facial-based DS identification. All images in the dataset were collected in compliance with ethical guidelines and anonymized to protect subject's privacy. The dataset consists of frontal facial images captured under various lighting conditions and backgrounds, representing a balanced distribution between DS and non-DS categories.

To ensure dataset quality and suitability for deep learning analysis, a set of explicit filtering criteria was applied. First, only images with a minimum resolution of at least 224 × 224 pixels were retained to preserve sufficient facial detail for feature extraction. All images were then resized to match the input requirements of each model, namely 224 × 224 pixels for MobileNetV3-Large and 240 × 240 pixels for EfficientNet-B1.

Images exhibiting excessive blur were excluded by visual inspection, supported by edge clarity assessment; images with indistinguishable facial contours or low sharpness were removed. In terms of illumination, images with extreme overexposure or underexposure where key facial regions such as the eyes, nose, or mouth were not clearly visible were discarded.

Furthermore, only frontal or near-frontal facial poses were included, with approximate head rotation limited to ±15 degrees to ensure consistent feature representation. Images with occlusions covering critical facial regions (e.g., eyes, nose, or mouth), such as masks, hands, or accessories, were also excluded. Additionally, images with incomplete facial

structures or dominant background interference were removed to maintain focus on relevant facial features.

These criteria were applied consistently across the dataset to ensure data quality, reduce noise, and improve the reliability and reproducibility of the model training process. These examples demonstrate the application of the predefined quality criteria and provide visual confirmation of the dataset consistency.

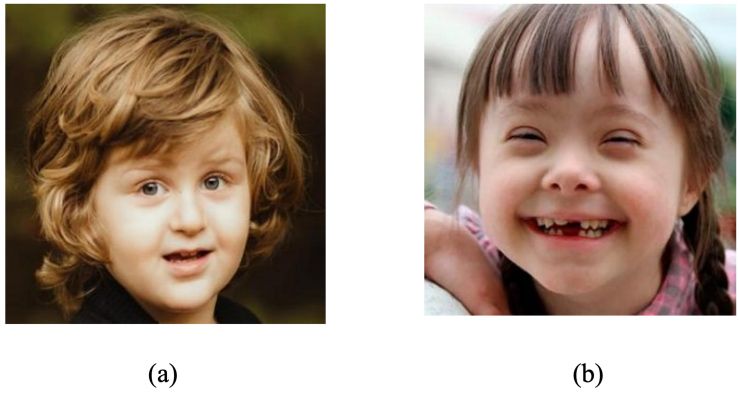

Figure 2: Sample image: (a) Non-DS, (b) DS.

Figure 2 presents representative examples of the facial image dataset after the quality filtering process. The images illustrate the visual differences between Non-DS and DS categories, highlighting distinctive facial characteristics that can be learned by the CNN models. These examples also reflect the quality criteria applied in this study, including clear facial visibility, frontal pose, and sufficient resolution.

2.2 Pre-processing Data

Figure 3 illustrates the dataset preparation process used in this study. The process began with the collection of 6,332 facial images as the initial dataset (raw dataset). The data then underwent filtering and quality control to remove low-quality images, such as those with blurriness, inadequate lighting, or inappropriate facial poses. After this stage, the data size was reduced to 3,030 images (the curated dataset).

In the next stage, the dataset underwent preprocessing and further refinement, resulting in a final dataset of 2,620 images used for model development. This process ensured that only high-quality and relevant facial images were used for model training and evaluation. After obtaining the final dataset, the data was divided into three subsets: training, validation, and testing data, with a ratio of 70:20:10. This division was carried out to ensure an optimal training process and fair and unbiased model evaluation.

The next stage was data augmentation, applied only to the training data. Unlike common augmentation approaches such as rotation or flipping, this study employed domain-based image enhancement techniques. Gaussian noise was added to approximately 23% of the training data to simulate variations in sensor conditions and improve the model's robustness to data variations [22].

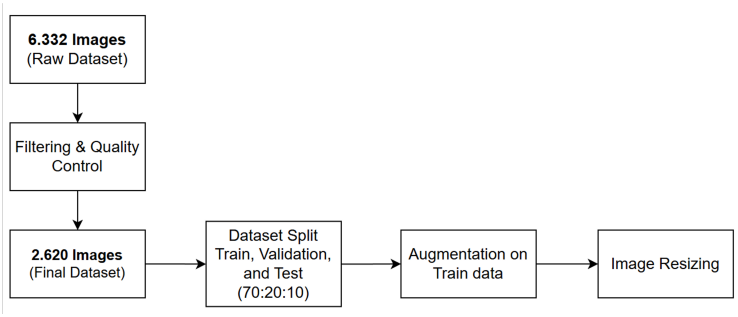

Figure 3: Dataset preparation and preprocessing workflow, illustrating the transition from raw data collection to the final dataset used for model training.

Furthermore, image sharpening using a convolution-based filter was applied to enhance edge details of facial features, and contrast adjustment was performed to address lighting variations in the images [23, 24]. These techniques aim to improve the model's ability to distinguish features without altering the underlying facial structure. The final stage is image resizing, during which all images are resized to the input size of each model architecture. Images for MobileNetV3-Large were resized to 224 × 224 pixels, while those for EfficientNet-B1 were resized to 240 × 240 pixels, in accordance with the standard specifications of each model.

Overall, this process was designed to ensure optimal data quality, improve model robustness to image variations, and support the model's performance and generalization capabilities in Down Syndrome classification. The final dataset distribution is shown in Table 1.

Table 1: Dataset distribution for training, validation, and testing

Dataset	Down Syndrome	Non-DS	Total
Training	916	916	1,832
Validation	262	262	524
Testing	132	132	264

Table 1 presents the final distribution of the dataset across training, validation, and testing subsets. The dataset is evenly divided between the Down Syndrome (DS) and Non-DS classes in each subset, ensuring a balanced representation for model training and evaluation. Specifically, the training set consists of 1,832 images, with 916 per class. The validation set contains 524 images, evenly distributed across 262 per class, while the testing set includes 264 images, with 132 per class. This balanced distribution is designed to minimize classification bias and to ensure that the model learns representative features from both categories [25].

2.3 Model Architecture

This research uses lightweight convolutional neural network (CNN) architectures: MobileNetV3-Large and EfficientNet-B1. These models perform binary classification of

facial images into DS and non-DS categories. Their efficiency in image classification and balance of performance led to their selection. This makes them suitable for use in environments with limited resources [17,26–28].

2.3.1 MobileNetV3-Large

MobileNetV3-Large is a compact CNN architecture specifically designed for effective inference on devices with limited computational capabilities. It achieves this by integrating depthwise separable convolutions (which use a single filter per input channel before merging the outputs), squeeze-and-excitation (SE) blocks (which dynamically adjust channel-wise feature responses), and the hard swish activation function (a non-linear function that enhances learning efficiency). Consequently, MobileNetV3-Large reduces computational expenses while ensuring robust feature extraction. This enables it to effectively learn distinctive facial features pertinent to DS classification, all while minimizing model size and inference latency. In this research, facial images are resized to 224 x 224 pixels to conform to the architecture's standard input requirements [27,28], as depicted in Figure 4.

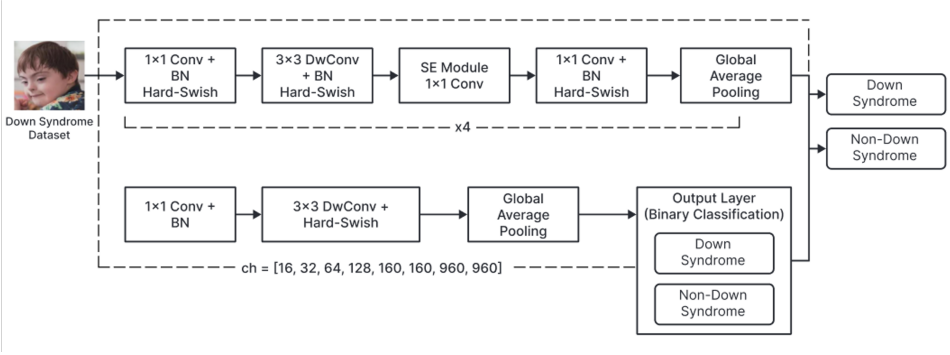

Figure 4: Architecture of the MobileNetV3-Large [29].

As illustrated in Figure 4, the input facial image (224 $\times$ 224 pixels) first passes through an initial convolutional layer, followed by batch normalization and the hard-swish activation function. The feature maps are then processed through a sequence of inverted residual bottleneck blocks, which consist of depthwise separable convolution and squeeze-and-excitation (SE) modules to enhance channel-wise feature representation [29].

These blocks are utilized repeatedly to incrementally derive more advanced facial features pertinent to DS classification. Subsequent to feature extraction, the resulting output is directed through a global average pooling, which diminishes spatial dimensions while retaining critical information. Ultimately, the refined features are input into a fully connected layer for binary classification, yielding the final output as either DS or Non-DS.

2.3.2 EfficientNet-B1

EfficientNet-B1 is a CNN architecture that employs a compound scaling approach to simultaneously adjust the network's depth, width, and input resolution harmoniously [17,26]. Compared with traditional CNN models, EfficientNet improves accuracy while using fewer parameters by effectively optimizing the distribution of computational resources.

EfficientNet-B1 is notably proficient in recognizing intricate visual patterns, which are crucial for facial analysis tasks where minor morphological variations are significant [17]. Consequently, EfficientNet-B1 is incorporated into this research to assess its ability to identify facial features associated with DS. The facial images processed with EfficientNet-B1 are resized to 240 x 240 pixels, which aligns with the architecture's recommended input resolution. The structure of the proposed model is depicted in Figure 5.

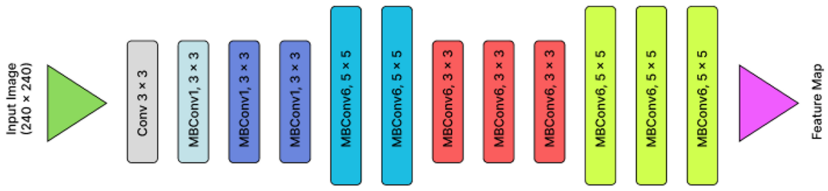

Figure 5: Architecture of the EfficientNet-B1 [26].

As shown in Figure 5, the input image (240 x 240 pixels) is first processed through an initial convolutional layer, followed by series of mobile inverted bottleneck convolution (MBConv) blocks. Each block integrated depthwise convolution, a squeeze-and-excitation optimization, and residual connections to efficiently capture both low-level and high-level facial features [17].

The architecture employs compound scaling, balancing depth, width, and resolution simultaneously to enhance feature representation. As the data propagates through successive MBConv layers, increasingly complex patterns are learned. The final feature maps are then aggregated using global average pooling and passed to a fully connected layer to generate the classification output. Compared to MobileNetV3-Large, EfficientNet-B1 emphasizes balanced scaling across network dimensions, leading to improved stability and generalization performance.

2.4 Model Training and Validation

The classification models in this study were developed using two lightweight convolutional neural networks architectures, namely MobileNetV3 and EfficientNet-B1. These architectures were selected due to their proven ability to achieve high classification performance while maintaining computational efficiency, which is essential for practical deployment in resource-constrained environments [26, 27]. MobileNetV3-Large is specifically designed to operate efficiently on devices with limited processing capabilities [27,28], whereas EfficientNet-B1 applies a compound scaling strategy to improve feature representation without significantly increasing computational cost [26].

All facial images used for training and evaluation were preprocessed according to the procedures described in Section 2.2. The curated dataset was divided into training, validation, and testing subsets using a stratified 70:20:10 ratio to ensure balanced class distributions across all partitions. This data partitioning strategy was adopted to obtain reliable performance estimation and to minimize potential bias during model training and evaluation.

During the training phase, both models were initialized with ImageNet-pretrained weights and fine-tuned using the binary cross-entropy loss function. The Adam optimizer

with a learning rate of 0.0001 was used to achieve stable, efficient convergence [30]. Model checkpoints were used to save the best-performing weights based on validation loss, ensuring that the final evaluation was conducted with the optimal model configuration.

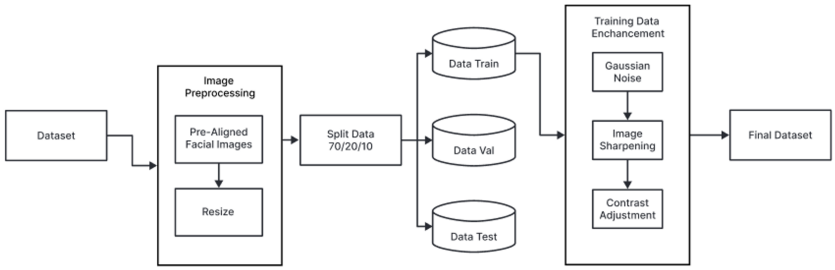

Figure 6: Preprocessing data flowchart.

Figure 6 illustrates the preprocessing pipeline applied in this study, including dataset preparation, image resizing, data splitting, and enhancement techniques such as Gaussian noise, image sharpening, and contrast adjustment. This pipeline standardizes and enriches input data before model training, improving robustness and generalization to variation in image quality and lighting.

Table 2: Model training hyperparameters

Parameter	**MobileNetV3**	**EfficientNet-B1**
Architecture	MobileNetV3-Large	EfficientNet-B1
Loss Function	Binary Cross-Entropy	Binary Cross-Entropy
Optimizer	Adam	Adam
Epoch	50	50
Learning Rate	0.0001	0.0001
Batch Size	16	16
Input Size	224×224	240×240

Table 2 presents a summary of the training hyperparameters employed for the MobileNetV3-Large and EfficientNet-B1 models. The uniform configuration across both architectures facilitates a fair and controlled comparison, whereas variations in input size correspond to the specific architectural needs of each model. The ultimate model was chosen based on achieving the lowest validation loss and all relevant training metrics were recorded for evaluation.

2.5 Model Evaluation

The effectiveness of each classification model was assessed using a confusion matrix, which provides a comprehensive overview of the model's ability to differentiate between DS and non-DS classes. From this confusion matrix, four frequently utilized metrics in binary classification were derived: accuracy, precision, recall, and F1-score, which are defined as follows [31]:

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

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

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

$$\text{F1Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

In this context, TP, TN, FP, and FN represent True Positive, True Negative, False Positive, and False Negative, respectively. These metrics comprehensively assess classification accuracy, which is crucial in healthcare applications where misclassification can have serious impact [31]. To ensure reliable and impartial performance assessment, K-fold cross-validation was used, aligning with contemporary deep learning research methodologies [19].

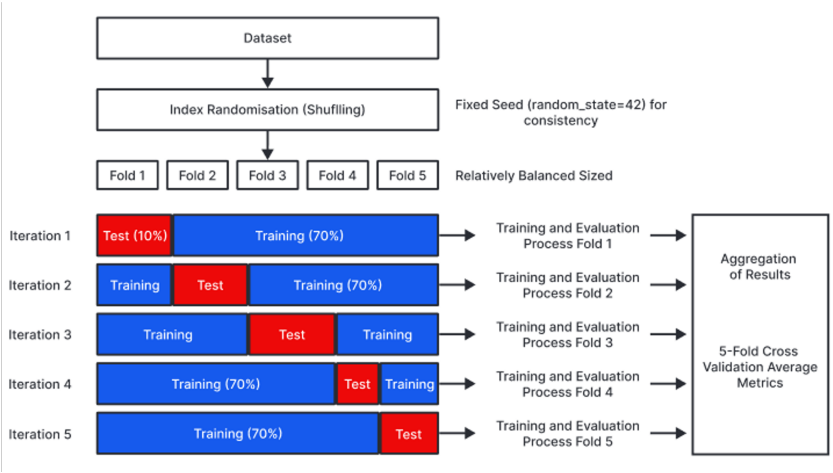

Figure 7: Evaluation workflow of the 5-fold cross-validation strategy with a 70:20:10 dataset split.

As shown in Figure 7, the dataset was split into training (70%), validation (20%), and test (10%) subsets. Cross-validation was done only on the training data, which was randomly shuffled and divided into 5 similar-sized folds. At each step, one fold served as internal validation and the rest for training. This was repeated until each fold was used as the validation set once. The model's final performance was evaluated using consolidated metrics across all folds.

Figure 8 provides additional insight into the evaluation phase that follows the cross-validation procedure. In contrast to Figure 7, which emphasizes data partitioning and the training process, this figure underscores the calculation of evaluation metrics for each fold

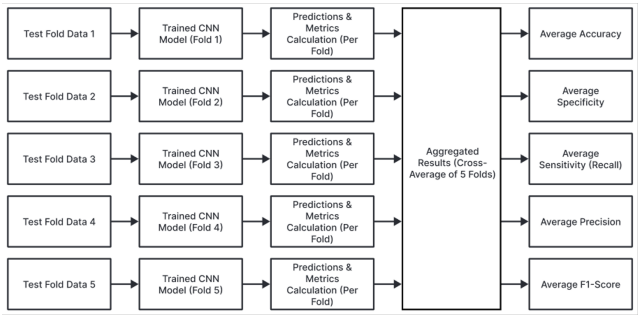

Figure 8: Model evaluation process across folds, including prediction, metric computation, and result aggregation.

and their subsequent aggregation to derive overall performance metrics. The average validation performance across all folds is calculated in the following manner:

$$\bar{M} = \frac{1}{K} \sum_{k=1}^{K} M_k \tag{5}$$

Where $\boldsymbol{M_k}$ represents the validation performance obtained from the k-th fold and $K = 5$ in this study. In addition to cross-validation, one-way Analysis of Variance (ANOVA) was applied to statistically assess whether the performance variations observed across folds were significant. The mean validation accuracy for each fold is defined as:

$$\bar{X}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij} \tag{6}$$

Where X_{ij} denotes the validation accuracy obtained from the i-th observation in the j-*th* fold, and n_j represents the number of observations in that fold. The overall mean across all folds is calculated as:

$$\bar{X} = \frac{1}{N} \sum_{j=1}^{K} \sum_{i=1}^{n_j} X_{ij} \tag{7}$$

The variability between folds is measured using the Sum of Squares between Groups (SSB):

$$SSB = \sum_{j=1}^{K} n_j (\bar{X}_j - \bar{X})^2 \tag{8}$$

While the variability within folds is measured using the Sum of Squares within Groups (SSW):

$$SSW = \sum_{j=1}^{K} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2 \tag{9}$$

The F-statistic is then computed as:

$$F = \frac{SSB/(k-1)}{SSW/(N-k)} \tag{10}$$

The null hypothesis H_0 assumes that there is no statistically significant difference among in mean validation performance across folds, whereas the alternative hypothesis H_1 assumes that at least one fold exhibits a statistically significant difference. A significance level of $p < 0.05$ was used to determine statistical significance, consistent with prior deep learning evaluation studies.

2.6 Model Interpretability Using Grad-CAM

To improve interpretability and understand the mechanism underlying each model's predictions, Grad-CAM was applied to MobileNetV3-Large and EfficientNet-B1. Grad-CAM functions by examining the gradients that pass into the last convolutional layer of a CNN, thereby facilitating the visualization of the spatial areas that have most significant impact on a specific class prediction [18].

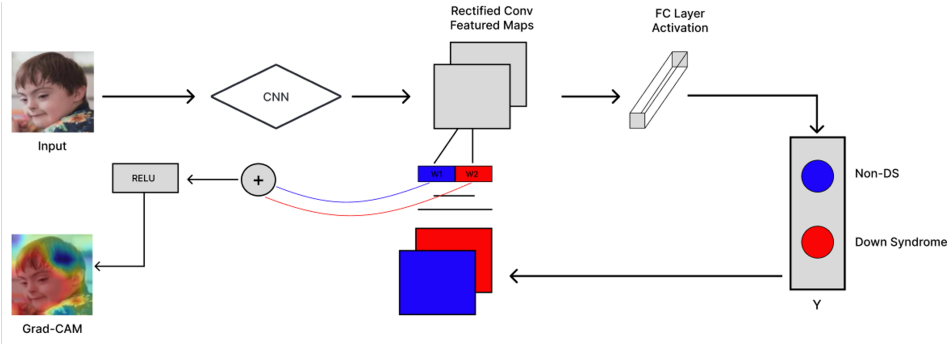

Figure 9: The Grad-CAM process for visualizing important regions in DS classification [32].

Figure 9 illustrates the Grad-CAM process for visualizing important regions in Down Syndrome (DS) classification. The process begins by feeding an input image into the CNN to obtain a class prediction. Subsequently, the gradient of the predicted class score with respect to the feature maps of the final convolutional layer is computed. These gradients are then globally averaged to produce channel-wise weights, which indicate the importance of each feature map for the target class.

The weighted feature maps are then combined to generate a coarse localization map. A Rectified Linear Unit (ReLU) activation function is applied to retain only the positive contributions, ensuring that the resulting heatmap highlights regions that positively influence the model's decision [33]. Mathematically, the Grad-CAM localization map for class c can be expressed as:

$$\mathrm{L^cGrad-CAM} = \mathrm{RELU}\left(\sum_k a_k^c A^k\right) \tag{11}$$

After obtaining the Grad-CAM formulation, the components of this expression can be understood as follows. The term $\mathrm{L^cGrad-CAM}$ represents the localization map for the

target class, showing the regions of the image that contribute most strongly to the model's prediction. The coefficient a_k^c denotes the importance weight of the k-th feature map and is computed as the global average of the gradients of the class score with respect to that feature map. Meanwhile A^k, it corresponds to the activation map produced by the k-th convolutional feature map in the final convolutional layer. Together, these components determine how each spatial location influences the output, and the ReLU function ensures that only positive contributions, those that support the predicted class, are retained in the final heatmap [18, 33, 34].

2.7 Comparative Analysis

This stage compares the performance of MobileNetV3-Large and EfficientNet-B1 using the evaluation metrics described in Section 2.5, including accuracy, precision, recall, and F1-score, K-Fold, and ANOVA. Both models underwent training and testing under the same conditions to guarantee a fair and uniform evaluation. In addition to quantitative metrics, Grad-CAM visualizations were examined to determine how each architecture emphasizes significant facial areas during prediction, offering further understanding of model interpretability. Computational factors, including inference speed and parameter efficiency, were also taken into account to evaluate the practical applicability of each model for real-world implementation.

3 Results

This section presents the experimental findings obtained from training and evaluating the EfficientNet-B1 and MobileNetV3-Large models for Down Syndrome (DS) facial image classification. The analysis includes training behavior, model performance comparison, cross-validation results, statistical testing, and Grad-CAM-based interpretability.

3.1 Training and Validation Results of EfficientNet-B1

The EfficientNet-B1 model was trained using a 5-fold cross-validation framework to evaluate its learning behavior and generalization capability across multiple data partitions. During the early stages of training, the model exhibited rapid convergence, with training accuracy exceeding 85% within the first few epochs. This indicates that the model efficiently learned discriminative facial features relevant to Down Syndrome (DS) classification.

Validation accuracy showed fluctuations during the initial epochs, reflecting an adaptation phase as the model adjusted to variations in facial characteristics and image quality. As training progressed, validation performance stabilized and gradually improved, reaching values above 90% in later epochs. Meanwhile, training loss decreased consistently across all folds, approaching near-zero values, while validation loss followed a generally decreasing trend, though it occasionally spiked.

Figure 10 presents the training and validation accuracy and loss curves of the EfficientNet-B1 model over 50 epochs. In the accuracy plot (left), training accuracy increases rapidly and approaches near-perfect performance, while validation accuracy shows early fluctuations before stabilizing at a high level. In the loss plot (right), training loss

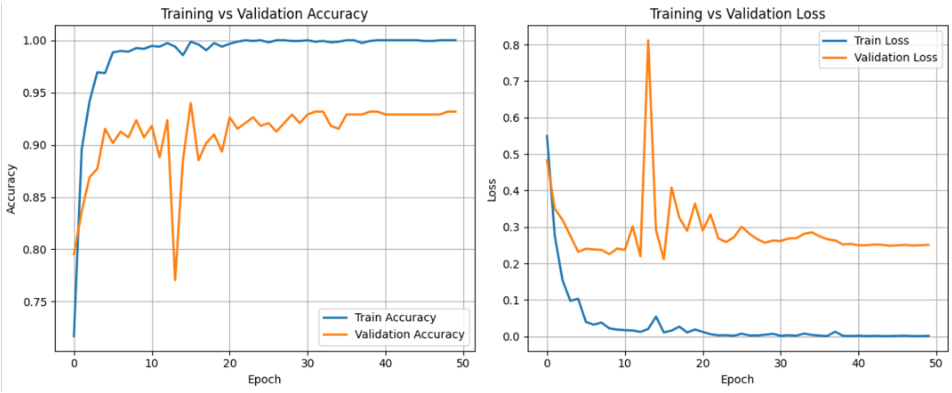

Figure 10: Training validation accuracy and loss curves of the EfficientNet-B1 model.

steadily decreases toward zero, indicating effective optimization. In contrast, validation loss shows an overall downward trend with several fluctuations, which may reflect data variability or the model's adaptation process during early training.

Across the five validation folds, EfficientNet-B1 demonstrates consistent performance with only minor variations in accuracy, indicating strong robustness to different training–validation splits. The stability of both accuracy and loss curves further confirms that the model converges smoothly and generalizes well across different data partitions.

Overall, these results indicate that EfficientNet-B1 achieves a well-balanced trade-off between learning capacity and generalization performance, with controlled overfitting and reliable predictive behavior, making it suitable for facial-based DS classification tasks.

3.2 Training and Validation Results of MobileNetV3-Large

The MobileNetV3-Large model was also evaluated using 5-fold cross-validation. Similar to EfficientNet-B1, MobileNetV3-Large exhibited rapid learning during the early epochs, with training accuracy exceeding 80% by the second epoch and approaching near-perfect performance in later stages. However, validation accuracy showed greater fluctuations, particularly during the first 10 epochs, indicating greater sensitivity to variations in the training data. Although the ReduceLROnPlateau scheduler helped control overfitting, the gap between training and validation performance remained more pronounced than in EfficientNet-B1. Test accuracy across folds ranged from approximately 86% to 90%, reflecting reasonable but less stable generalization capability. Figure 11 presents the training and validation accuracy and loss curves for the MobileNetV3-Large model, illustrating relatively unstable convergence and greater variability between the training and validation trends. These findings demonstrate that while MobileNetV3-Large is computationally efficient and capable of learning discriminative features, its performance is more sensitive to data partitioning than that of EfficientNet-B1.

Figure 11 presents the training and validation accuracy and loss curves of the MobileNetV3-Large model over 50 epochs. In the accuracy plot (left), the training accuracy increases rapidly during the initial epochs and approaches near-perfect performance, indicating that the model effectively learns discriminative features from the training data.

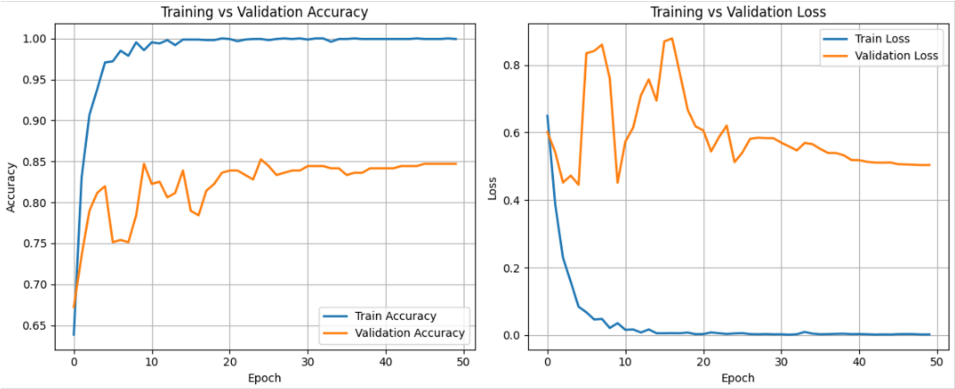

Figure 11: Training validation accuracy and loss curves of the MobileNetV3-Large model.

However, the validation accuracy exhibits noticeable fluctuations, particularly within the first 10 epochs, before gradually stabilizing at around 84–85%.

In the loss plot (right), the training loss decreases sharply and approaches near-zero values, indicating efficient optimization. In contrast, the validation loss shows significant fluctuations and remains relatively higher throughout the training process, with several spikes observed in the early and middle epochs. This pattern suggests that the model is more sensitive to variations in the validation data and may experience mild overfitting.

Overall, the divergence between the training and validation curves indicates that although MobileNetV3-Large is capable of learning useful features, its generalization performance is less stable than that of EfficientNet-B1. The observed variability across epochs reflects greater sensitivity to data partitioning, which may compromise the consistency of the model's predictions.

3.3 Performance Comparison

A detailed performance comparison between the two models was conducted using confusion matrix analysis. EfficientNet-B1 correctly classified 119 DS and 122 Non-DS samples, with 13 false negatives and 10 false positives. In contrast, MobileNetV3-Large correctly identified 100 DS and 128 Non-DS samples but produced a substantially higher number of false negatives (32 cases). Figure 12, illustrates the confusion matrix of the EfficientNet-B1 model, highlighting its balanced classification performance between DS and Non-DS classes.

This imbalance indicates that MobileNetV3-Large tends to favor Non-DS predictions, resulting in reduced sensitivity for DS detection. From a medical screening perspective, this is a critical limitation, as false negatives may delay diagnosis and early intervention. EfficientNet-B1, on the other hand, demonstrated more balanced classification across both classes, making it more suitable for real world DS screening applications.

Following the analysis of EfficientNet-B1, the performance of MobileNetV3-Large was also examined using confusion matrix analysis. Figure 13 presents the confusion matrix of the MobileNetV3-Large model. In contrast to EfficientNet-B1, MobileNetV3-Large correctly classified 100 DS samples and 128 Non-DS samples. However, the number of DS samples misclassified as Non-DS increased significantly to 32 false negatives, while only 4

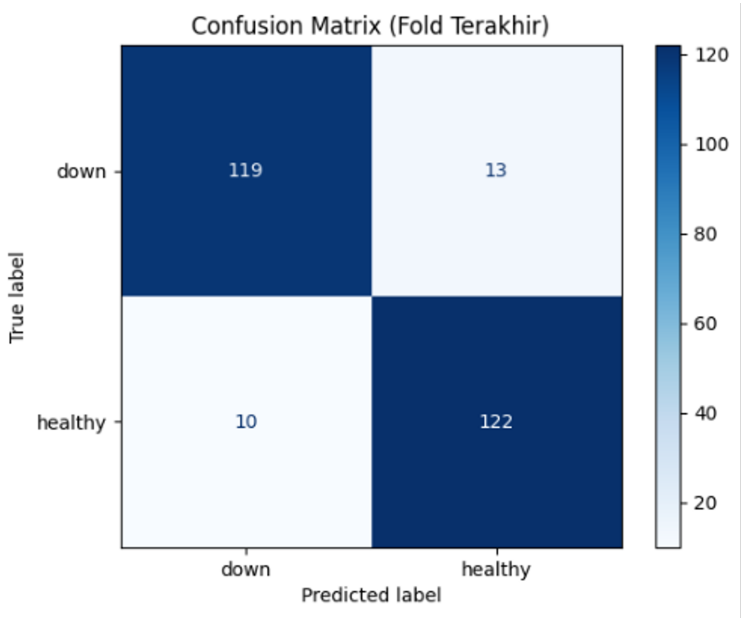

Figure 12: Confusion matrix of EfficientNet-B1 performance.

Non-DS samples were misclassified as DS. This pattern indicates that MobileNetV3-Large exhibits a stronger bias toward predicting the Non-DS class, resulting in reduced sensitivity for DS identification. From a clinical screening perspective, such a bias is problematic, as a high false negative rate may delay diagnosis and early intervention.

3.4 K-Fold Cross-Validation and Statistical Analysis

To ensure robust evaluation, both models were assessed using 5-fold cross-validation. Table 3 summarizes the cross-validation performance of EfficientNet-B1 and MobileNetV3-Large across all folds. EfficientNet-B1 achieved test accuracies ranging from 89.39% to 92.80%, demonstrating relatively stable performance across different data partitions. In contrast, MobileNetV3-Large exhibited greater variability, with test accuracies between 86.36% and 90.15%, indicating higher sensitivity to changes in training–validation splits. One-way ANOVA was performed on epoch-wise validation accuracies to examine statistical differences among folds. For both models, p-values were significantly below 0.05, indicating that data partitioning had a measurable influence on model performance. Nevertheless, as reflected in Table 3, EfficientNet-B1 consistently showed smaller performance variations compared to MobileNetV3-Large, confirming its superior stability and robustness for facial-based DS classification tasks.

To further determine whether the observed differences between the two models were statistically significant, a one-way ANOVA analysis was conducted on the validation accuracies obtained from each fold. The outcomes of this statistical test are summarized in Table 4. The ANOVA results confirm that for both models the p-values were below 0.05, indicating a significant effect of data partitioning on model performance. However, when

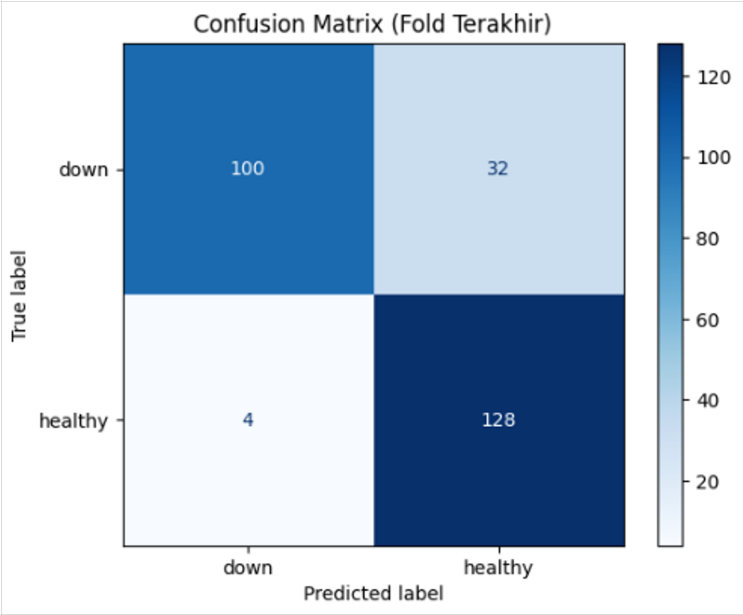

Figure 13: Confusion matrix of MobileNetV3-Large performance.

Table 3: K-Fold performance of EfficientNet-B1

Fold	Test Accuracy	Test Loss	Last Validation Accuracy
1	0.9280	0.2950	0.9537
2	0.9167	0.3302	0.9046
3	0.8939	0.4512	0.9208
4	0.9053	0.3478	0.9044
5	0.9129	0.3272	0.9317

comparing the two architectures, EfficientNet-B1 demonstrated lower variability and more consistent behavior, reinforcing the performance trends presented in Table 3.

Table 4: K-Fold performance of MobileNetV3-Large

Fold	Test Accuracy	Test Loss	Last Validation Accuracy
1	0.8977	0.3852	0.9264
2	0.9015	0.3833	0.8665
3	0.8788	0.5192	0.8907
4	0.8864	0.4203	0.8962
5	0.8636	0.5496	0.8470

Together, Table 3 and Table 4 provide complementary evidence: Table 3 illustrates the empirical performance stability of each model, while Table 4 statistically validates the observed differences, confirming the superior robustness of EfficientNet-B1 for facial based DS classification.

3.5 Grad-CAM Results

Grad-CAM visualization was applied to analyze the interpretability of both models and to identify the facial regions that contributed most to the classification decisions. Figure 14 illustrates the Grad-CAM visualization results for DS classification using EfficientNet-B1 (a) and MobileNetV3-Large (b), presenting the original input images, generated heatmaps, and overlay representations. The visualization demonstrates that EfficientNet-B1 generated more focused and localized activation maps, primarily concentrated on clinically relevant regions such as the forehead, periocular area, and midface. In contrast, MobileNetV3-Large produced broader and less concentrated attention patterns, indicating less discriminative feature localization.

These findings suggest that EfficientNet-B1 learns more reliable and meaningful facial features, thereby enhancing confidence in its decision-making process. The Grad-CAM analysis confirms that both models base their predictions on relevant facial characteristics rather than on irrelevant background information. However, the more precise and consistent attention patterns produced by EfficientNet-B1 indicate superior interpretability, which is essential for the practical deployment of AI-based screening tools in medical environments.

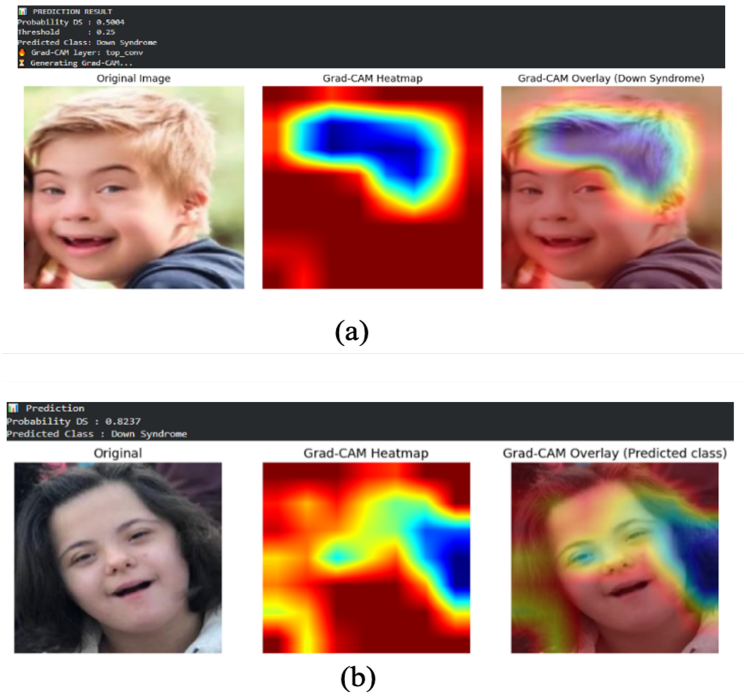

Figure 14: Grad-CAM visualization for DS classification using EfficientNet-B1 (a) and MobileNetV3-Large (b), showing the original image, heatmap, and overlay.

4 Discussion

This study aimed to evaluate the effectiveness, robustness, and interpretability of two lightweight CNN architectures EfficientNet-B1 and MobileNetV3-Large, for facial-based DS classification.

4.1 Performance Analysis

The experimental findings indicate that both models achieved strong classification performance. However, EfficientNet-B1 consistently demonstrated higher accuracy, better stability across folds, and more balanced sensitivity between DS and Non-DS classes. MobileNetV3-Large, while efficient, showed greater variability and a higher false-negative rate, which is less desirable in medical screening contexts.

The application of 5-fold cross-validation and ANOVA confirmed that model performance is influenced by data partitioning, emphasizing the importance of rigorous validation in medical AI research. EfficientNet-B1 maintained superior average performance despite these variations, indicating stronger generalization capability.

4.2 Model Interpretability Using Grad-CAM

Interpretability analysis revealed that EfficientNet-B1 produced more localized and anatomically consistent attention maps than MobileNetV3-Large. This suggests that EfficientNet-B1 relies on more relevant facial features when making predictions, increasing its suitability for clinical decision support.

4.3 Comparison with Previous Research

The accuracy achieved by EfficientNet-B1 (91.14%) aligns with the performance range reported in the comprehensive review by [9], which indicates that DCNN-based methods typically achieve 90–98% accuracy. Although some prior studies reported higher accuracy, many relied on single train–test splits without robust validation. In contrast, this study applied cross-validation, statistical testing, and interpretability analysis, providing a more reliable and transparent evaluation framework.

4.4 Theoretical and Practical Implications

Beyond performance comparison, this study provides important contributions to both the theoretical understanding and practical implementation of deep learning models for Down Syndrome (DS) classification.

From a theoretical perspective, this study advances current knowledge by demonstrating that model robustness and stability are critical factors alongside accuracy in medical image classification. While many previous studies emphasize achieving high accuracy, this work shows that performance consistency across different data splits can significantly affect a model's reliability. The use of cross-validation and ANOVA further underscores the need to incorporate statistical analysis to ensure reproducibility and avoid misleading conclusions from single train–test splits.

In addition, this study contributes to the growing body of research on lightweight CNN architectures by providing a structured comparative analysis between MobileNetV3-Large and EfficientNet-B1 specifically for DS facial classification. Unlike prior work, which often evaluates models independently, this comparison offers deeper insight into the trade-offs among accuracy, stability, and computational efficiency, thereby helping guide model selection in future research.

From a practical perspective, the findings of this study support the development of deployable and trustworthy AI-based screening systems. The superior stability and interpretability of EfficientNet-B1 suggest it is better suited for real-world medical applications, particularly in resource-limited settings where both computational efficiency and decision transparency are essential.

Furthermore, integrating Grad-CAM enhances model interpretability by providing visual explanations of the decision-making process, thereby improving clinician trust and facilitating the adoption of AI systems in healthcare environments. Therefore, this study not only compares model performance but also provides actionable insights for selecting reliable and interpretable models for DS screening applications.

4.5 Limitations of the Study

Despite promising results, several limitations exist. The dataset size and demographic diversity were limited, which may restrict generalizability. The models were evaluated on retrospective data and have not yet been validated in real clinical environments. Additionally, only two architectures were investigated.

4.6 Future Work

Future research should focus on multi-center data collection, exploration of additional architectures, and prospective clinical validation. Integration into mobile-based screening tools is also an important direction.

5 Conclusion

This study presents a comparative analysis of two lightweight convolutional neural network architectures, EfficientNet-B1 and MobileNetV3-Large, focusing on face-based DS classification performance. Both models demonstrated strong classification accuracy, indicating the effectiveness of deep learning for DS screening using facial images. Between the two, EfficientNet-B1 consistently outperformed MobileNetV3-Large in terms of validation stability and balance of classification outcomes across folds, demonstrating greater robustness and reliability under varied evaluation conditions.

To further support these results, K-fold cross-validation and one-way ANOVA confirmed the reliability of the evaluation process, revealing statistically significant performance differences across folds for both models. Together, these findings highlight the importance of cross-validation and statistical analysis in medical image classification to mitigate bias and ensure reproducibility. In addition, Grad-CAM visualization improves model interpretability by highlighting facial regions associated with DS characteristics, thereby increasing transparency and supporting clinical confidence in the model's decision-making

process. Looking ahead, future research should focus on collecting multi-center and multi-ethnic datasets to improve generalizability. Additional deep learning architectures and ensemble approaches can also be explored. Furthermore, integrating the model into mobile-based screening applications and conducting clinical validation with medical experts are important directions for further development.

6 Acknowdledge

This research was supported by the Institute for Research and Community Service (LPPM), Universitas Syiah Kuala, and conducted in collaboration with Universiti Sains Islam Malaysia (USIM). The authors sincerely appreciate the institutional support provided by both institutions.

References

[1] W. K. Chung and F. F. Herrera, "Health supervision for children and adolescents with syndrome," *Pediatrics*, vol. 9, no. 4, 2023.

[2] K. Windsperger and S. Hoehl, "Development of down syndrome research over the last decades–what healthcare and education professionals need to know," *Frontiers in Psychiatry*, vol. 12, pp. 1–7, 2021.

[3] 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 central java, indonesia," *Intractable Rare Diseases Research*, vol. 13, no. 2, pp. 121–125, 2024.

[4] E. Zevanya, W. Indrarto, D. Lestari, and T. M. M. Widagdo, "Maternal age increases the risk of down syndrome: A case-control study in yogyakarta, indonesia," *Berkala Ilmiah Kedokteran Duta Wacana*, vol. 9, no. 1, 2024.

[5] T. R. Simamora, S. Y. Patria, and S. Wandita, "Congenital heart disease, gastrointestinal defect, and low birth weight as the contributing factors for three-year survival rates among down syndrome children in indonesia," *Indonesia Journal of Biomedical Science*, vol. 16, no. 2, pp. 65–69, 2022.

[6] B. Qin, L. Liang, J. Wu, Q. Quan, Z. Wang, and D. Li, "Automatic identification of down syndrome using facial images with deep convolutional neural network," *Diagnostics*, vol. 10, no. 7, p. 487, 2020.

[7] E. Setyati, S. Az, S. P. Hudiono, and F. Kurniawan, "Cnn based face recognition system for patients with down and william syndrome," *Knowledge Engineering and Data Science*, vol. 4, no. 2, pp. 138–144, 2021.

[8] E. Elyan, P. Vuttipittayamongkol, P. Johnston, K. Martin, K. McPherson, C. F. Moreno-García, C. Jayne, and M. M. K. Sarker, "Computer vision and machine learning for medical image analysis: recent advances, challenges, and way forward," *Artificial Intelligence Surgery*, 2022.

[9] K. Rezaee, "Machine learning and facial recognition for down syndrome detection: A comprehensive review," *Computers in Human Behavior Reports*, vol. 17, p. 100600, 2025.

[10] M. A. Shaikh, H. S. Al-Rawashdeh, and A. R. W. Sait, "Deep learning-powered down syndrome detection using facial images," *Life*, vol. 15, p. 1361, 2025.

[11] M. C. Thomas and S. P. Arjunan, "Deep learning measurement model to segment the nuchal translucency region for early identification of down syndrome," *Measurement Science Review*, vol. 22, no. 4, pp. 187–192, 2022.

[12] Yunidar, Roslidar, Yusni, Nasaruddin, and FitriArnia, "Cnn performances for stunting face image classification," in *International Conference on Electrical Engineering and Computer Science*, 2024, pp. 89–94.

[13] I. Ramadhan, M. Melinda, Y. Yunidar, D. D. Acula, R. Miftahujjannah, S. Rusdiana, and Z. Zainal, "Mobile application development for facial classification of autistic children based on mobilenet-v3," *Jurnal Infotel*, vol. 17, no. 3, pp. 612–626, 2025.

[14] 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 cnn," *Radioelectronics and Computer Systems*, no. 1, pp. 76–86, 2024.

[15] Y. Yunidar, Y. Yusni, N. Nasaruddin, F. Arnia *et al.*, "Cnn performance improvement for classifying stunted facial images using early stopping approach," *Jurnal RESTI*, vol. 9, no. 1, pp. 62–68, 2025.

[16] M. Irhamsyah, M. Melinda, Y. Yunidar, I. Muttaqin, and L. Q. Zakaria, "Implementation of dwt and xception for ecg image classification of arrhythmic patients," *Jurnal Infotel*, vol. 17, no. 2, pp. 336–356, 2025.

[17] P. Upadhyay, "Face recognition using efficientnet," in *International Conference on Intelligent Applications*, 2023, pp. 7–14.

[18] T. A. Araf, A. Siddika, S. Karimi, and M. G. R. Alam, "Real-time face emotion recognition and visualization using grad-cam," in *International Conference on Advanced Electrical, Computer, Communication and Sustainable Technologies*, 2022, pp. 1–5.

[19] I. K. Nti, O. Nyarko-Boateng, J. Aning *et al.*, "Performance of machine learning algorithms with different k values in k-fold cross validation," *International Journal of Information Technology and Computer Science*, vol. 13, no. 6, pp. 61–71, 2021.

[20] S. Elbassuoni, H. Ghattas, J. El Ati, Z. Shmayssani, S. Katerji, Y. Zoughbi, A. Semaan, C. Akl, H. B. Gharbia, and S. Sassi, "Deepnova: A deep learning nova classifier for food images," *IEEE Access*, vol. 10, pp. 128 523–128 535, 2022.

[21] "Detection of down syndrome computer vision project," https://universe.roboflow.com/projects-qbm9c/detection-of-down-syndrome, 2024.

[22] R. S. Thakur, S. Chatterjee, R. N. Yadav, and L. Gupta, "Image de-noising with machine learning: A review," *IEEE Access*, vol. 9, pp. 93 338–93 363, 2021.

[23] M. G. Kumar and A. D. Goswami, "Automatic classification of the severity of knee osteoarthritis using cnn," *Applied Sciences*, vol. 13, no. 3, 2023.

[24] A. Rofena, V. Guarrasi, M. Sarli, C. L. Piccolo, M. Sammarra, B. B. Zobel, and P. Soda, "A deep learning approach for virtual contrast enhancement in spectral mammography," *Computerized Medical Imaging and Graphics*, vol. 116, p. 102398, 2024.

[25] K. Ghosh, C. Bellinger, R. Corizzo, P. Branco, B. Krawczyk, and N. Japkowicz, "The class imbalance problem in deep learning," *Machine Learning*, vol. 113, no. 7, 2024.

[26] W. Hastomo, A. S. B. Karno, E. Sestri, V. Terisia, D. Yusuf, S. A. Arman, and D. Arif, "Classification of brain tumor images using efficientnet," *Semesta Teknika*, vol. 27, no. 1, pp. 46–54, 2024.

[27] F. Zhu, Y. Sun, Y. Zhang, W. Zhang, and J. Qi, "An improved mobilenetv3 mushroom quality classification model using images with complex backgrounds," *Agronomy*, vol. 13, no. 12, 2023.

[28] M. Abd Elaziz, A. Dahou, N. A. Alsaleh, A. H. Elsheikh, A. I. Saba, and M. Ahmadein, "Boosting covid-19 image classification using mobilenetv3 and aquila optimizer algorithm," *Entropy*, vol. 23, no. 11, 2021.

[29] A. Alsenan, B. Ben Youssef, and H. Alhichri, "Mobileunetv3—a combined unet and mobilenetv3 architecture for spinal cord gray matter segmentation," *Electronics*, vol. 11, no. 15, 2022.

[30] M. Reyad, A. M. Sarhan, and M. Arafa, "A modified adam algorithm for deep neural network optimization," *Neural Computing and Applications*, vol. 35, no. 23, pp. 17 095–17 112, 2023.

[31] S. Sathyanarayanan and B. R. Tantri, "Confusion matrix-based performance evaluation metrics," *African Journal of Biomedical Research*, vol. 27, no. 4S, pp. 4023–4031, 2024.

[32] H. Moujahid, B. Cherradi, M. Al-Sarem, L. Bahatti, A. B. A. M. Y. Eljialy, A. Alsaeedi, and F. Saeed, "Combining cnn and grad-cam for covid-19 disease prediction and visual explanation." *Intelligent Automation & Soft Computing*, vol. 32, no. 2, 2022.

[33] H. Panwar, P. Gupta, M. K. Siddiqui, R. Morales-Menendez, P. Bhardwaj, and V. Singh, "A deep learning and grad-cam based color visualization approach for fast detection of covid-19 cases using chest x-ray and ct-scan images," *Chaos, Solitons & Fractals*, vol. 140, p. 110190, 2020.

[34] S. Guluwadi *et al.*, "Enhancing brain tumor detection in mri images through explainable ai using grad-cam with resnet 50," *BMC medical imaging*, vol. 24, no. 1, pp. 1–19, 2024.