VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7890

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

STUNTING CLASSIFICATION IN CHILDREN USING VIOLA-JONES AND MULTI-FEATURE FUSION WITH PRE-TRAINED MODELS

Maylani Kusuma Wardhani[1]; Garin Muhammad Akbar[1]; Christian Sri Kusuma Aditya[1*]

Department of Informatics [1]
Universitas Muhammadiyah Malang, Indonesia [1]
https://umm.ac.id[1]
maylaniwardhani@webmail.umm.ac.id, muhakbar12@webmail.umm.ac.id, christianskaditya@umm.ac.id*

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *Stunting remains a critical public health issue, particularly in developing countries, where early detection plays a vital role in prevention and intervention. Previous studies have generally relied on single-feature approaches, either using handcrafted descriptors or convolutional neural networks (CNNs) alone, which often fail to capture subtle craniofacial differences associated with stunting. This study proposes an image-based classification system for detecting stunting in children using facial analysis. The proposed method integrates Viola–Jones face detection with facial landmarks, Gray Level Co-occurrence Matrix (GLCM), Color Co-occurrence Matrix (CCM), and local descriptors such as SIFT–FAST/ORB, combined with deep features extracted from a pre-trained EfficientNet model. Feature fusion was performed by concatenating handcrafted and deep features before classification using a fully connected layer with Softmax activation. Experimental results demonstrated that the proposed fusion model achieved superior performance compared to single-feature baselines, reaching 98% accuracy, 0.98 precision, 0.97 recall, and an F1-score of 0.98. These findings indicate that the integration of geometric, texture, color, and deep semantic cues effectively enhances sensitivity toward the stunting class and improves model interpretability. The novelty of this study lies in the combination of classical computer vision and deep learning techniques for robust, interpretable, and clinically relevant stunting detection. This approach offers strong potential for developing digital health tools that enable early, non-invasive stunting screening in children.*

Keywords: *EfficientNet, GLCM, Image Landmark, Stunting Classification, Viola-Jones*

Intisari— *Stunting masih menjadi masalah kesehatan masyarakat yang krusial, terutama di negara berkembang, di mana deteksi dini berperan penting dalam upaya pencegahan dan intervensi. Penelitian terdahulu umumnya menggunakan pendekatan fitur tunggal, baik berbasis deskriptor buatan maupun jaringan saraf konvolusional (CNN), yang sering kali gagal mengenali perbedaan halus pada struktur wajah yang berkaitan dengan stunting. Penelitian ini mengusulkan sistem klasifikasi berbasis citra untuk mendeteksi stunting pada anak melalui analisis wajah. Metode yang diusulkan mengintegrasikan deteksi wajah Viola–Jones dengan penanda wajah (facial landmarks), Gray Level Co-occurrence Matrix (GLCM), Color Co-occurrence Matrix (CCM), serta deskriptor lokal seperti SIFT dan FAST/ORB, yang kemudian dikombinasikan dengan fitur dalam (deep features) dari model EfficientNet pra-latih. Fusi fitur dilakukan dengan menggabungkan fitur buatan dan fitur dalam sebelum diklasifikasikan menggunakan jaringan saraf berlapis penuh dengan aktivasi Softmax. Hasil eksperimen menunjukkan bahwa model fusi yang diusulkan mencapai kinerja tertinggi dengan akurasi 98%, presisi 0,98, recall 0,97, dan skor F1 sebesar 0,98. Integrasi fitur geometrik, tekstur, warna, dan semantik mendalam terbukti meningkatkan sensitivitas terhadap kelas stunting dan memperkuat interpretabilitas model. Pendekatan ini memiliki potensi besar untuk diterapkan pada sistem kesehatan digital sebagai alat deteksi dini stunting yang non-invasif dan relevan secara klinis.*

Kata Kunci: *EfficientNet, GLCM, Landmark Citra, Klasifikasi Stunting, Viola-Jones.*

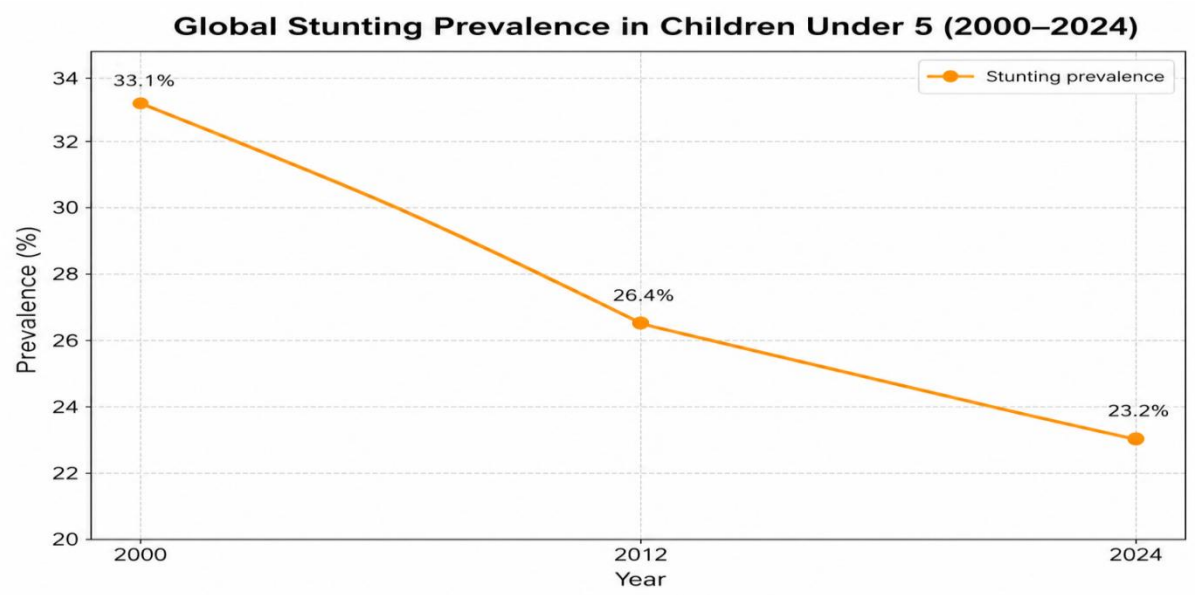
.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7890

INTRODUCTION

Stunting is a major global public health problem that affects over 150 million children worldwide, particularly in low- and middle-income countries [1]. It is defined as impaired linear growth resulting from chronic malnutrition, recurrent infections, and inadequate feeding practices during early life [2]. Children suffering from stunting not only experience delays in physical growth but are also at higher risk of cognitive impairment, reduced productivity, and increased susceptibility to chronic diseases in adulthood [3]. Figure 1 presents Global prevalence of stunting among children under five (2000–2024).



Source: (UNICEF/WHO/World Bank Joint Child Malnutrition Estimates [4], [5], 2024)
Figure 1. Global Prevalence of Stunting Among Children Under Five (2000–2024)

Global stunting prevalence has declined over the past decade but progress has recently stagnated. UNICEF reported a decrease from 26.4% in 2012 to 23.2% in 2024, while the World Bank noted a similar plateau in regions affected by poverty and malnutrition [4], [6]. These trends indicate that current efforts remain insufficient, underscoring the need for innovative approaches to early stunting detection.

Recent studies indicate that facial morphology can serve as a reliable indicator of developmental status, as proportions such as inter-eye distance, nose length, and jawline shape are correlated with stunting [7], [8], [9]. This finding supports non-invasive, image-based detection approaches that use facial features to assess growth conditions, particularly in resource-limited settings.

However, previous studies still present several limitations. Most existing approaches rely on single-feature representations, such as geometric or texture-based features, which may not fully capture the complex and subtle craniofacial variations associated with stunting. Furthermore, some studies are constrained by relatively small datasets and lack comprehensive evaluation using multiple performance metrics, which may affect the robustness and generalization capability of their findings. Therefore, a more comprehensive approach that integrates multiple feature representations is required to better capture discriminative information and improve classification performance and reliability.

Advances in computer vision have made such methods feasible through algorithms like Viola–Jones for face detection and handcrafted descriptors such as GLCM and CCM for capturing texture and color information [3], [10], [11]. Local descriptors including SIFT and FAST/ORB further enhance robustness to variations in lighting and pose, while prior studies confirm their effectiveness in identifying subtle facial characteristics linked to growth abnormalities [1].

Deep learning methods, particularly pre-trained convolutional neural networks such as EfficientNet, continue to demonstrate strong performance in medical image analysis due to their balanced scaling of depth, width, and resolution [12], [13], [14]. Evidence from recent studies further confirms that integrating deep features

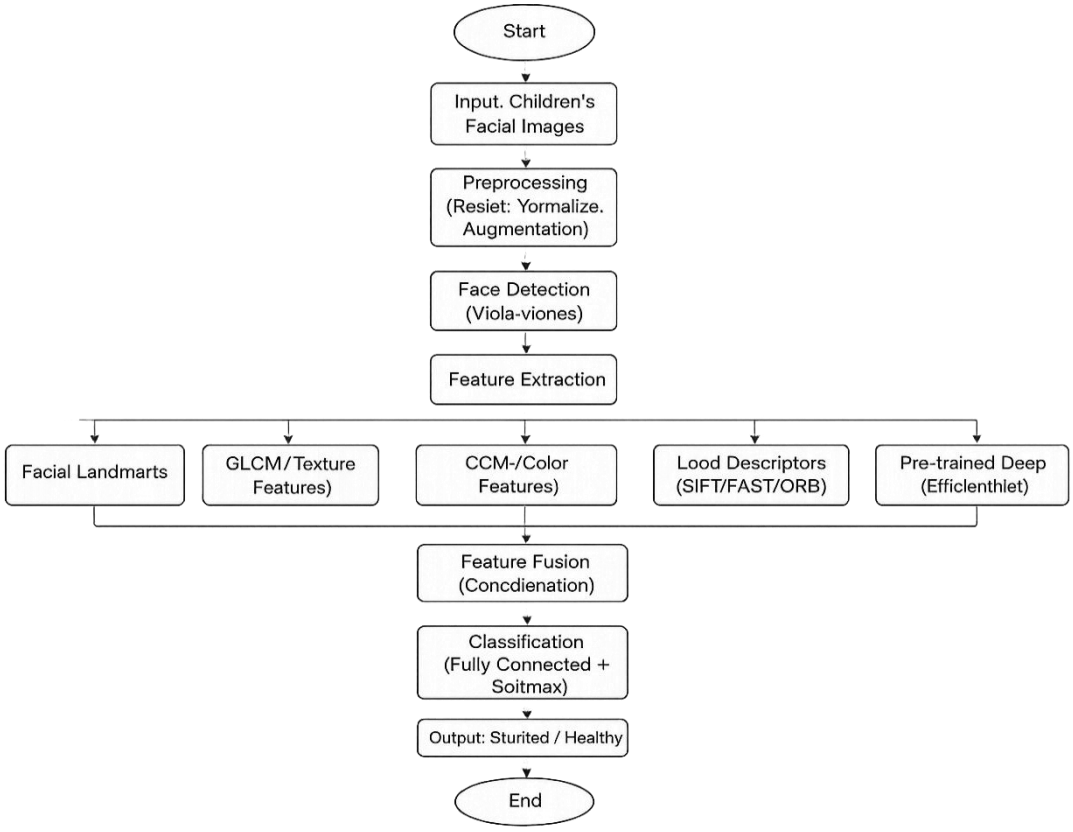
with handcrafted descriptors can increase a model's sensitivity in capturing subtle structural variations within biological images [15]. Research on stunting detection also shows the effectiveness of neural network–based approaches, including Multi-Layer Perceptron models that successfully classify stunting with high evaluation metrics, indicating the relevance of deep feature extraction in this domain [16]. A similar trend is found in medical imaging, where EfficientNet-based architectures consistently outperform prior CNN models in classification accuracy, reinforcing their suitability for tasks involving complex visual patterns such as craniofacial analysis [17]. Building on these advancements, this study introduces a multi-feature fusion framework that combines Viola–Jones detection, craniofacial landmarks, GLCM, CCM, local descriptors (SIFT/FAST), and EfficientNet deep features to enhance representation quality.

However, existing studies still face several limitations. Most approaches rely on single or partially integrated feature representations, which may not fully capture the complex interactions between structural, textural, and semantic facial characteristics. In addition, many methods primarily emphasize accuracy without explicitly addressing sensitivity (recall), which is crucial in medical-related classification tasks such as stunting detection. Furthermore, model interpretability is often overlooked, limiting the ability to verify whether the model focuses on clinically relevant facial regions.

To address these gaps, this study proposes a comprehensive framework that systematically integrates multiple complementary feature representations while emphasizing balanced performance, particularly in improving recall to minimize false negatives. In addition, Grad-CAM visualization is incorporated to enhance model interpretability. This integrated approach provides a more robust, reliable, and clinically meaningful solution compared to existing methods.

MATERIALS AND METHODS

This study employed a structured methodology encompassing dataset acquisition, preprocessing, face detection, feature extraction, feature fusion, model training, and evaluation as illustrated in Figure 2.



Source: (Research Results, 2025)

Figure 2. Overall Stunting Detection Methodology Flowchart System

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7890

Dataset and Preprocessing

The original dataset consisted of 1,577 children's facial images collected from three publicly available datasets hosted on Roboflow [18], [19], [20]. These datasets were curated to ensure anonymity and ethical compliance. The images from the three sources were merged into a single dataset to increase diversity and improve model generalization. The images were preprocessed and annotated using the Roboflow platform. Each image was labeled as stunted or healthy based on standardized anthropometric indicators defined by UNICEF and WHO growth standards. To enhance model robustness and mitigate overfitting, data augmentation techniques including rotation, horizontal flipping, scaling, and brightness variation were applied, increasing the effective training dataset to approximately 3,000 images. Demographic metadata such as geographic origin, ethnicity, and socioeconomic background were not available in the source datasets. [3], [21].

The dataset was divided into training, validation, and testing sets using an 80:10:10 ratio. To prevent data leakage, data augmentation was applied exclusively to the training set after the splitting process, ensuring that no augmented samples were included in the validation or test sets. Due to the absence of subject identity information in the source datasets, subject-wise splitting could not be conducted. This limitation is acknowledged, as it may introduce potential bias and affect the generalization performance due to possible intra-subject similarities across data splits.

The dataset was nearly balanced, with 55% healthy and 45% stunted children, covering age groups between 2 and 10 years. Images varied in framing (full-body, upper-body, and close-up facial shots), background conditions, and illumination. Stunted children typically showed craniofacial differences such as narrower jawlines, reduced facial symmetry, and smaller inter-orbital distance, consistent with medical studies [1]. Since the dataset included images with heterogeneous framing, preprocessing was required to ensure consistency:

a) Cropping and isolation of the face region. Full-body and upper-body images were cropped so that only the face remained.
b) Background removal. Hair, clothing, and irrelevant background were excluded.
c) Resizing and normalization. All cropped face images were resized to 224 × 224 pixels to conform to the input requirements of the EfficientNet architecture employed in this study, ensuring compatibility and optimal computational performance. Furthermore, pixel values were normalized to the range [0,1] to improve training stability, accelerate convergence, and reduce numerical instability during model optimization.
d) Data augmentation. Rotation (±15°), horizontal flipping, and brightness variation were applied to improve robustness.

Source: (Research Results, 2025)

Figure 3. Examples of Raw Images and Cropped Faces (Healthy vs. Stunted)

Figure 3 presents Examples of raw images and cropped faces (healthy vs. stunted).The dataset was divided into 80% training, 10% testing, and 10% validation subsets to ensure balanced learning and fair evaluation. As shown in Figure 3, the dataset contains cropped facial images of both healthy and stunted children. Medical literature notes that stunted children often exhibit narrower inter-eye distance, a more prominent forehead, smaller jaw structure, and mild facial asymmetry [1]. These visual traits highlight structural differences relevant for early detection of growth impairments.

Face Detection with Viola–Jones

The Viola–Jones algorithm was employed to automate and standardize facial region detection [13]. It utilizes Haar-like features with an AdaBoost cascade classifier and sliding-window search to ensure accurate localization of frontal faces. Each detected face was enclosed in a bounding box adjusted to include the forehead and chin, capturing key growth-related regions. From these detected faces, landmark points such as inter-eye distance, nose length, mouth-to-chin ratio, and facial symmetry were extracted to represent geometric features correlated with stunting [15]. Figure 4 shows examples of the detected facial regions and corresponding geometric mappings.

Source: (Research Results, 2025)

Figure 4. Viola–Jones Face Detection With Bounding Boxes and Facial Landmark Mesh.

Feature Extraction

Feature extraction aimed to capture multi-dimensional facial characteristics through geometric (Viola–Jones and landmarks), textural (GLCM), color-based (CCM), local (SIFT/FAST/ORB), and deep semantic (EfficientNet) features. This combination integrated handcrafted and deep learning descriptors to recognize both structural and visual cues distinguishing stunted from healthy children. Handcrafted features captured spatial and chromatic details, while deep embeddings provided higher-level semantic representations, resulting in improved robustness, accuracy, and interpretability in stunting classification [7], [8], [1], [15], [22].

a. Facial Landmarks

Facial landmarks were extracted using a 68-point model that localized key anatomical points on each child's face, including the eyes, eyebrows, nose, mouth, and jawline. These landmarks represent the geometric structure of the face and were used to compute ratios such as inter-eye distance, nose length, and mandibular width. Such geometric proportions have been found to correlate with craniofacial developmental indicators in medical literature [1], [23]. Normalization was performed by dividing each ratio by the overall bounding box dimensions, ensuring that results were invariant to image scale and camera distance. This step reduced geometric bias and allowed meaningful comparison across samples. By leveraging landmark-based measurements, this method provided structural insight complementary to texture and color-based descriptors [14], [24].

b. Gray-Level Co-occurrence Matrix (GLCM)

The Gray-Level Co-occurrence Matrix (GLCM) was applied to analyze spatial relationships between pixel intensities and capture local texture patterns associated with structural characteristics [7], [8], [25]. Four key descriptors contrast, energy, homogeneity, and correlation were computed to represent intensity variation, texture uniformity, smoothness, and regularity. These texture features provided valuable cues for distinguishing facial surface differences between stunted and healthy children [1], [26]. Texture features were derived from GLCM [25]. Four descriptors were computed:

$$Contrast = \sum_{i,j}(i-j)^2\ P(i,j) \quad (1)$$

$$Energy = \sum_{i,j}\ P(i-j)^2 \quad (2)$$

$$Homogeneity = \sum_{i,j}\ \frac{P\ (i,j)}{1+|i-j|} \quad (3)$$

$$Correlation = \sum_{i,j}\ \frac{(i-\mu_i)\ (j-\mu_i)\ P\ (i,j)}{\sigma_i \sigma_j} \quad (4)$$

Contrast, as shown in Equation (1), is used to measure the degree of intensity variation between neighboring pixels, which reflects the roughness and local texture irregularities in an image. Energy, defined in Equation (2), measures the uniformity of pixel distribution, where higher values indicate more structured and homogeneous textures. Homogeneity, presented in Equation (3), calculates the proximity of pixel intensity values to the GLCM diagonal, indicating smoother and more consistent texture surfaces. Meanwhile,

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7890

Correlation, in Equation (4), evaluates the linear relationship between pixel pairs, capturing directional patterns and texture regularity within the image.

c. Color Co-occurrence Matrix (CCM)

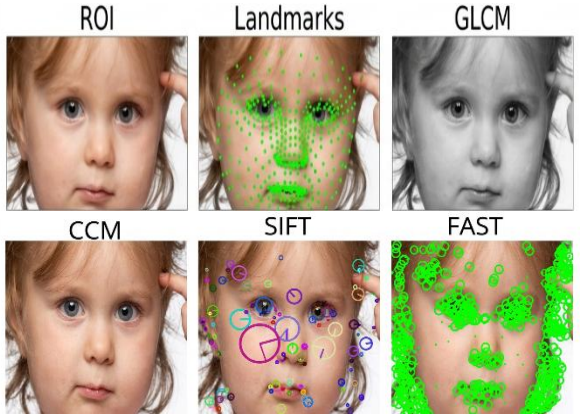
Color-based features were extracted using the Color Co-occurrence Matrix (CCM), which analyzes relationships between color intensities across the RGB channels. This descriptor captures pigmentation patterns and chromatic texture that may be associated with nutritional or growth status. In this study, six CCM descriptors were extracted: mean, variance, contrast, energy, entropy, and correlation [26], [27]. Mean and variance represented the average and dispersion of color intensities. Contrast reflected chromatic differences between neighboring pixels, while energy described the uniformity of color patterns. Entropy measured color randomness or disorder, and correlation quantified the dependency between color channels. These descriptors provided a rich representation of skin color distribution and tone uniformity, complementing structural and textural information captured by GLCM and geometric features [1], [28].

d. Local Descriptors (SIFT, FAST/ORB)

Local descriptors such as SIFT, FAST, and ORB were applied to capture stable keypoints and local gradient patterns in facial images that remain consistent under variations in scale, rotation, and illumination [9], [27]. These algorithms effectively identified fine geometric and textural details such as edges, contours, and corners enhancing the model's ability to detect subtle craniofacial asymmetries. By encoding these localized features, the descriptors improved the discriminative power and sensitivity of stunting classification [22], [29].

e. Pre-trained EfficientNet

EfficientNet-B0, a convolutional neural network pre-trained on ImageNet, was utilized to extract deep semantic features. This architecture applies compound scaling to balance network depth, width, and resolution, thereby achieving high accuracy with fewer parameters [11], [30]. The extracted 1280-dimensional feature vectors captured abstract and high-level visual information that complemented handcrafted features. When combined with geometric, texture, and color descriptors, these deep embeddings contributed to a more comprehensive and robust facial representation. Figure 5 illustrates examples of extracted features, including GLCM texture maps, CCM color distributions, SIFT/FAST keypoints, and EfficientNet Grad-CAM visualizations [22], [29]. Figure 5 presents Feature extraction results.



Source: (Research Results, 2025)

Figure 5. Feature Extraction Results

Feature Fusion, Model Training, and Evaluation

All extracted features geometric (Viola–Jones and landmarks), textural (GLCM), color-based (CCM), local (SIFT/FAST/ORB), and deep semantic (EfficientNet) were concatenated into a unified vector to integrate handcrafted and deep learning representations for better discrimination between stunted and healthy children. The fusion vector was processed through a fully connected neural network with ReLU activation and Softmax output, implemented using TensorFlow and Keras [11], [22], [31].

The model was trained using the Adam optimizer with a learning rate of 0.001, a batch size of 32, and 50 epochs. The learning rate of 0.001 was adopted due to its widespread use and empirically demonstrated ability to achieve stable and efficient convergence across various deep learning tasks. A batch size of 32 was employed as a trade-off between computational efficiency and memory limitations, while still ensuring stable gradient updates. The training process was conducted for up to 50 epochs, with early stopping applied if no improvement was observed for 10 consecutive epochs to mitigate overfitting. These hyperparameters were determined based on commonly adopted practices and empirical experimentation, where the selected configuration resulted in stable convergence and satisfactory generalization performance.

However, it is important to note that these values may not represent globally optimal hyperparameters for all scenarios. The dataset was split into 80% training, 10% validation, and 10% testing. The model evaluation was conducted using a hold-out validation strategy supported by a dedicated validation set to monitor training and

prevent overfitting. While cross-validation could provide a more robust assessment, it was not implemented due to the computational complexity of the proposed multi-feature fusion framework and the relatively large augmented dataset size. This limitation is acknowledged and will be addressed in future work. Performance evaluation using Accuracy, Precision, Recall, and F1-score demonstrated strong classification performance, supported by confusion matrix analysis and Grad-CAM visualizations highlighting clinically relevant facial regions [3], [29], [32].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

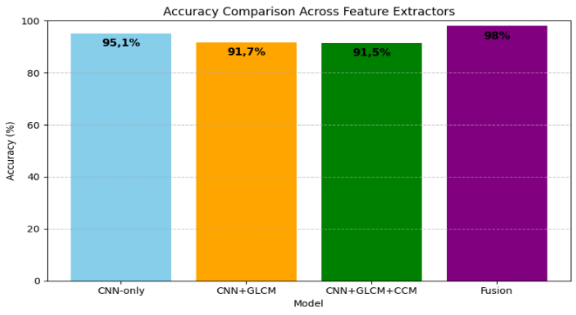
$$F1\text{-}Score = 2x\frac{Precision\ x\ Recall}{Precision + Recall} \quad (8)$$

Accuracy, Equation (5), is used to measure how often the model produces correct predictions overall. Precision, Equation (6), measures the model's ability to avoid generating false positive predictions. Recall, Equation (7), evaluates the model's capability to correctly detect all relevant positive cases. Meanwhile, F1-Score, Equation (8), represents the harmonic mean of precision and recall, providing a balanced evaluation between the two metrics.

Several prior studies have implemented Gray Level Co-occurrence Matrix (GLCM) for texture-based image analysis. Poetri et al. from Universitas Muhammadiyah Malang applied GLCM to classify facial skin types using contrast, energy, homogeneity, and entropy features [33]. Another study from UMM utilized GLCM combined with DBSCAN for texture-based image retrieval, demonstrating the robustness of GLCM descriptors in image clustering tasks [34].

RESULTS AND DISCUSSION

This study evaluated various feature combinations for classifying stunting in children. The baseline model using EfficientNetB3 achieved 95.1% accuracy, with a recall of 97.2%, indicating that although the model was capable of identifying most stunted cases, its overall performance was still limited due to imbalance in other evaluation metrics [22].



Source: (Research Results, 2025)

Figure 6. Performance Comparison of Feature Extraction Results

Figure 6 illustrates the performance comparison across all feature extraction configurations. The CNN-only baseline produces the lowest overall performance in terms of feature representation capability, indicating its limitation in capturing subtle craniofacial differences associated with stunting. A noticeable improvement appears when GLCM features are added, resulting in an accuracy of 91.7% and recall of 92.9%.

However, when CCM features are further integrated, the model achieves an accuracy of 91.5% with a recall of 91.4%. Although this configuration shows a slight decrease in both accuracy and recall compared to the previous stage, it still demonstrates the contribution of color-based features in capturing chromatic variations of facial characteristics. These results suggest that texture and color information complement deep features, even though their combined effect may not always produce a strictly monotonic performance increase [26].

The final fusion model combining CNN embeddings, GLCM, CCM, local descriptors (SIFT/FAST), and Viola–Jones geometric features achieved 98% classification accuracy, with 0.988 precision, 0.975 recall, and an F1-score of 0.981. This demonstrates that integrating geometric, textural, and semantic representations significantly improved sensitivity and balance in detecting stunted children [3].

Table 1. Performance and Computational Efficiency Comparison Across Feature Configurations

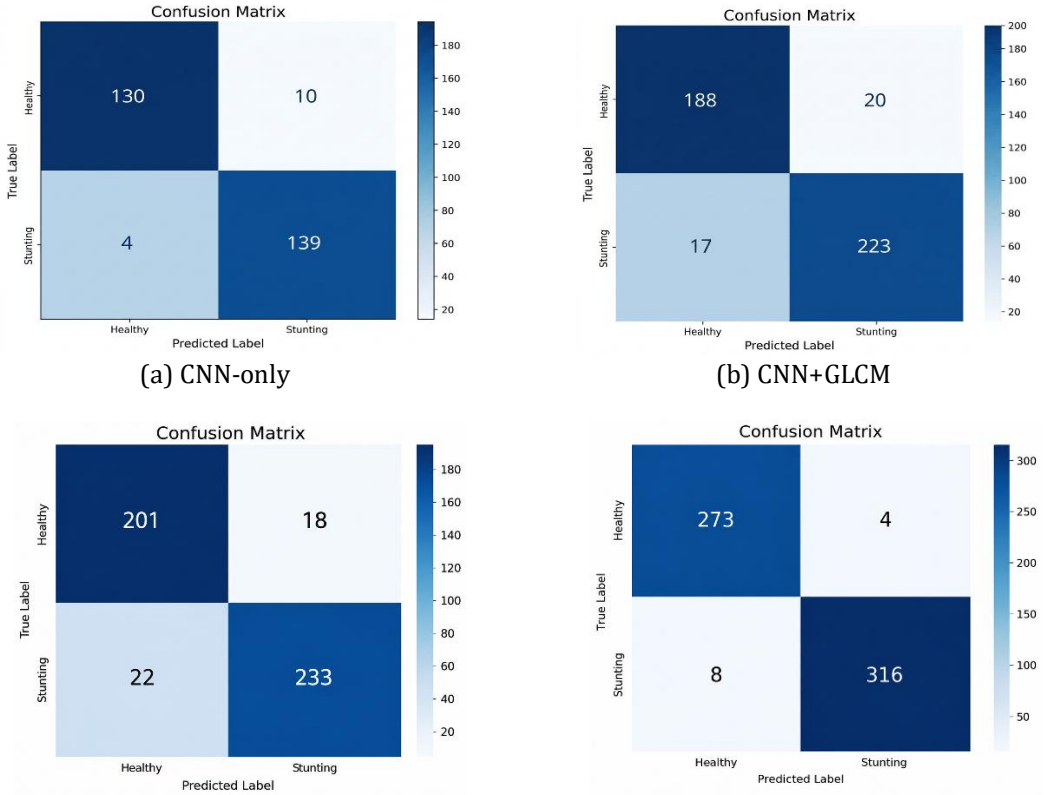
Configuration	Total Trainable Parameters	Model Size (MB)	Processing Time (ms/image)
Baseline CNN	11,577,136	133.26	93.59
CNN + GLCM	11,203,312	128.60	142.02
CNN + GLCM + CCM	11,204,080	128.61	66.47
CNN + Multi-Feature Fusion	11,404,016	130.90	66.09

Source: (Research Results, 2025)

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7890

To evaluate the computational feasibility of the proposed feature fusion strategy, a comparative analysis was conducted based on the number of trainable parameters, model size, and processing time across different feature configurations. As presented in Table 1, the multi-feature fusion model contains 11.40 million trainable parameters, which is comparable to the baseline CNN with 11.57 million trainable parameters, indicating only a marginal difference in model complexity. Furthermore, the proposed fusion model achieves a processing time of 66.09 ms per image, which is faster than the baseline CNN (93.59 ms) and substantially more efficient than the CNN + GLCM configuration. These results demonstrate that integrating multiple feature types improves representation capability while maintaining computational efficiency and avoiding significant additional overhead.



(a) CNN-only (b) CNN+GLCM

(c)CNN+GLCM+CCM (d) CNN+GLCM+CCM+Local+Viola–Jones

Source: (Research Results, 2025)

Figure 7. Confusion matrices for all feature extraction configurations: (a) CNN-only, (b) CNN+GLCM, (c) CNN+GLCM+CCM, and (d) CNN+GLCM+CCM+Local+Viola–Jones fusion model.

Figure 7 presents the confusion matrices for all experimental configurations. The results clearly indicate a progressive improvement from single-feature to multi-feature models. In particular, confusion matrices (c) and (d), corresponding to CNN+GLCM+CCM and CNN+GLCM+CCM+Local+Viola–Jones, exhibited significantly higher true positive and true negative counts compared to earlier baselines. The improvement in (c) occurred because the integration of GLCM and CCM features enabled the model to capture both textural and chromatic variations on the facial surface, such as skin tone uniformity and pigmentation differences, which CNN features alone could not represent [26].

Meanwhile, the highest values observed in (d) were achieved when local and geometric descriptors were added through SIFT/FAST and Viola–Jones. These additional features enhanced the model's ability to recognize structural cues such as inter-eye distance, facial width-to-height ratio, and jawline contour, all of which are clinically relevant to growth development. As a result, the fusion model in (d) demonstrated more balanced predictions and substantially reduced false negatives for the stunting class, indicating higher sensitivity and reliability in distinguishing facial variations associated with stunting [3], [29]. To further illustrate the use of evaluation metrics, an example calculation based on the confusion

matrix in Figure 7(d) is presented. In this study, the stunting class is considered as the positive class, with TP = 316, TN = 273, FP = 4, and FN = 8. Based on these values, the evaluation metrics are computed as follows:

Accuracy = (TP + TN) / (TP + TN + FP + FN) = (316 + 273) / 601 = 0.980

Precision = TP / (TP + FP) = 316 / 320 = 0.988

Recall = TP / (TP + FN) = 316 / 324 = 0.975

F1-score = 2 × (Precision × Recall) / (Precision + Recall) = 0.981

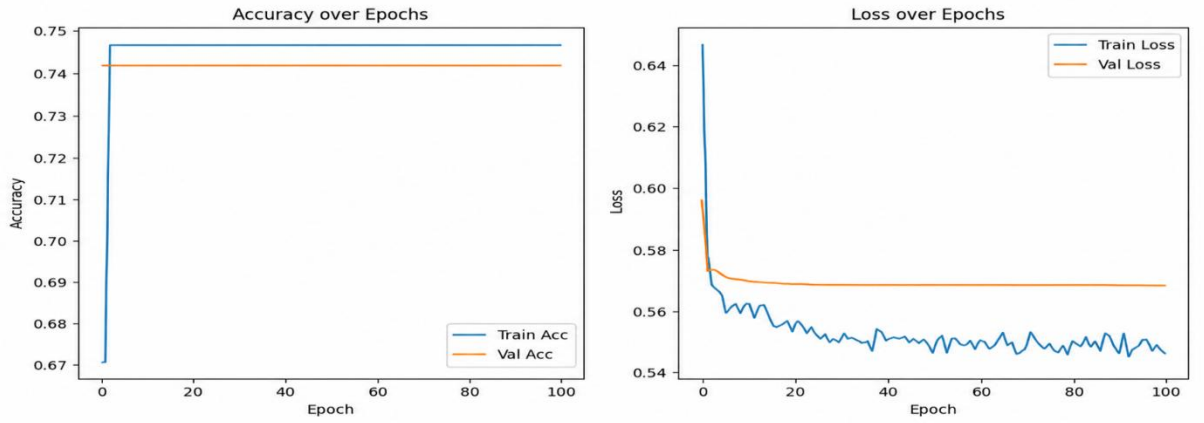
Furthermore, a comparison of recall values across all configurations is provided to support the claim of improved sensitivity. The recall values for models (a), (b), (c), and (d) are 0.972, 0.929, 0.914, and 0.975, respectively. These results indicate that although the baseline model (a) achieved relatively high recall, its overall performance remained limited in other evaluation metrics. In contrast, the proposed fusion model (d) not only achieved the highest recall but also maintained superior accuracy, precision, and F1-score, indicating a more balanced and robust classification performance. The high recall achieved by model (d) highlights its effectiveness in minimizing false negatives, which is particularly important in stunting detection to ensure that affected children are not overlooked.

Source: (Research Results, 2025)

Figure 8. Representative misclassification examples between Healthy and Stunting classes.

Figure 8 illustrates the misclassification results, demonstrating that the model encounters difficulty in accurately distinguishing between the Healthy and Stunting classes. This issue is primarily attributed to the high visual similarity between the two classes, as some stunted children do not exhibit sufficiently distinctive facial characteristics compared to healthy children, resulting in overlapping feature representations. Additionally, image distortions, inconsistent aspect ratios, and variations in pose, illumination, and background conditions may reduce the quality of the extracted features and limit the model's ability to capture discriminative patterns. Dataset-related limitations, such as limited diversity and potential class imbalance, may further affect the model's generalization performance. Consequently, these factors contribute to misclassification, indicating the importance of improving data quality, increasing dataset diversity, and enhancing feature extraction methods to achieve more robust and reliable classification performance.

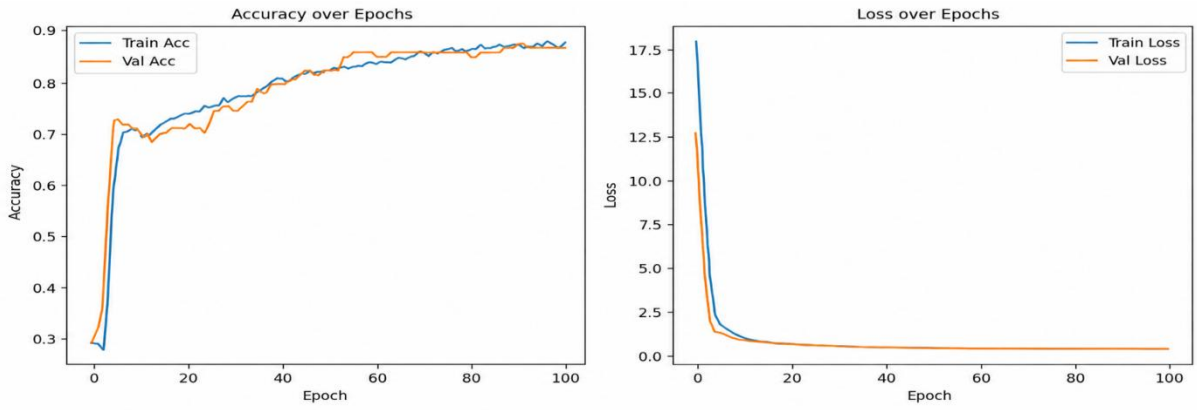
Figure 9 shows that the fusion model converged steadily with a minimal gap between training and validation accuracy, indicating stable learning behavior and no evident statistical overfitting. However, it is important to distinguish between conventional overfitting (train–validation divergence) and potential dataset-specific bias.



(a) CNN-only,

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7890



(b) Fusion model

Source: (Research Results, 2025)

Figure 9. (a)-(b) Training and validation performance curves: (a) CNN-only model and (b) fusion model integrating CNN, GLCM, CCM, local descriptors, and Viola–Jones features.

Figure 9 shows that the fusion model converged steadily with a minimal gap between training and validation accuracy, indicating stable learning behavior and no evident statistical overfitting. However, it is important to distinguish between conventional overfitting (train–validation divergence) and potential dataset-specific bias. To mitigate the risk of learning irrelevant cues such as background artifacts or illumination patterns, facial cropping, normalization, and augmentation techniques were applied during preprocessing. Furthermore, Grad-CAM visualizations demonstrated that model attention was predominantly concentrated on central facial structures rather than peripheral image regions. While these findings suggest that the model primarily relies on structural facial information, they do not completely eliminate the possibility of dataset-specific bias. Therefore, further validation using external and demographically diverse datasets is required to confirm broader generalizability.

Source: (Research Results, 2025)

Figure 10. Grad-CAM visualizations showing discriminative facial regions: forehead, orbital area, and jawline.

Figure 10 presents Grad-CAM visualizations showing discriminative facial regions: forehead, orbital area, and jawline. Furthermore, the Grad-CAM visualization played a crucial role in validating the interpretability of the proposed model. By generating heatmaps from the final

convolutional layers, Grad-CAM highlighted specific facial areas that contributed most to the classification decision. The results showed that the fusion model consistently focused on anatomically relevant regions such as the forehead, orbital area, nose bridge, and jawline features that are clinically correlated with craniofacial growth abnormalities in stunted children. These findings indicate that the model's attention was primarily concentrated on facial regions rather than peripheral background areas, but rather on meaningful structural indicators, thereby strengthening its clinical credibility and reliability for early stunting assessment [29].

While the highlighted regions correspond to anatomical areas associated with craniofacial development, it is important to clarify that Grad-CAM identifies regions of model attention rather than establishing direct clinical causality. The observed activation patterns may reflect morphological variations related to soft tissue distribution, facial proportionality, or skeletal growth differences commonly reported in malnutrition-related developmental delays. However, these visual explanations should be interpreted as supportive evidence of model focus rather than definitive biological markers. Further clinical and anthropometric validation studies are required to confirm the physiological relevance of these facial regions in stunting assessment.

Overall, the integration of handcrafted and deep features successfully enhanced both performance and reliability. The fusion strategy proved superior to single-feature methods by capturing diverse facial cues, leading to a model that is not only accurate but also interpretable in a medical context. This demonstrates strong potential for implementing computer-vision-based tools for early stunting detection, particularly in resource-limited healthcare environments.

The experimental results revealed that CNN-only models using EfficientNetB3 exhibited clear limitations, as they tended to overfit to the majority class and failed to identify subtle facial differences in stunted children. This weakness confirmed that deep features alone were insufficient to represent fine-grained craniofacial variations related to growth abnormalities. The inclusion of handcrafted descriptors, particularly the Gray-Level Co-occurrence Matrix (GLCM) and Color Co-occurrence Matrix (CCM), significantly enhanced the model's sensitivity by capturing texture and color patterns that reflected facial surface irregularities. These improvements demonstrated that handcrafted features contributed complementary information to deep embeddings, resulting in better representation of visual cues associated with stunting.

The most substantial improvement occurred with the multi-feature fusion model, which combined geometric, textural, color, and deep semantic features into a unified representation. This integration achieved balanced accuracy and high recall, confirming the robustness of the proposed approach in detecting growth abnormalities. The Grad-CAM visualization further strengthened the interpretability of the model by highlighting medically relevant facial regions such as the forehead, orbital area, and jawline, which align with clinical studies on craniofacial growth. These findings suggest that the proposed fusion framework not only improves classification performance but also supports early stunting detection by providing explainable and clinically reliable visual evidence [29], [31].

CONCLUSION

This study proposed a multi-feature fusion framework for classifying stunting in children by integrating Viola–Jones geometric features, facial landmarks, textural descriptors (GLCM), color descriptors (CCM), local keypoint features (SIFT/FAST), and deep semantic features extracted from a pre-trained EfficientNet model. The experimental results showed that single-feature approaches such as CNN-only models failed to effectively identify stunted children due to class imbalance and the inability to capture subtle craniofacial variations. The integration of handcrafted and deep learning features substantially improved classification performance. The proposed fusion model achieved the best results with an overall accuracy of 98%, precision of 0.98, recall of 0.97, and an F1-score of 0.98. These findings indicate that geometric, texture, color, and semantic cues complement one another, enabling the model to detect subtle facial differences associated with stunting more effectively.

Theoretically, this research suggests that combining classical computer vision methods with deep learning can enhance both performance and interpretability, addressing some limitations of single-approach techniques. From a practical perspective, the proposed framework shows potential as a non-invasive screening support tool. However, further validation using clinical datasets and standardized anthropometric measurements is required before it can be applied in real-world healthcare settings. In addition, the current evaluation is based on a hold-out validation

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 4 MAY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i4.7890

strategy, and future work incorporating cross-validation could provide a more comprehensive assessment of model robustness. Future research will prioritize structured multi-center data collection across diverse geographic regions and demographic groups to enable systematic fairness evaluation and bias assessment. Comprehensive demographic metadata annotation (e.g., age subgroup, sex, and regional background) will be incorporated to facilitate subgroup performance analysis and ensure equitable model behavior. To enhance deployment feasibility in low-resource settings, model optimization strategies such as pruning and quantization will be explored to reduce computational overhead while preserving predictive performance. In addition, robustness testing under real-world imaging conditions including variations in illumination, occlusion, and motion blur will be systematically conducted to evaluate practical reliability. Furthermore, prospective clinical and external validation studies comparing model predictions with standardized anthropometric measurements will be undertaken to establish medical credibility and ensure safe clinical applicability. In the long term, integration into a lightweight mobile-based screening application is envisioned to support community-level early stunting detection while maintaining ethical safeguards and human oversight.

REFERENCE

[1] N. A. Hidayat, D. S. Latif, and A. S. Setiawan, "Facial Proportions in Stunted and Non-Stunted Children Aged 7 – 72 Months : A Cross-Sectional Study in Bandung , Indonesia," *Children*, vol. 12, no. 8, p. 1037, 2025, doi: 10.3390/children12081037.

[2] H. Shen, H. Zhao, and Y. Jiang, "Machine Learning Algorithms for Predicting Stunting among Under-Five Children in Papua New Guinea," *Children*, vol. 10, no. 10, p. 1638, 2023, doi: 10.3390/children10101638.

[3] UNICEF, WHO, and World Bank Group, "Levels and Trends in Child Malnutrition: Key Findings of the 2023 Edition of the Joint Child Malnutrition Estimates," Geneva, 2023.

[4] N. H. Vu, N. M. Trieu, H. Nguyen, A. Tuan, and T. D. Khoa, "applied sciences Review : Facial Anthropometric , Landmark Extraction , and Nasal Reconstruction Technology," *Applied Sciences*, vol. 12, no. 19, p. 9548, 2022, doi: 10.3390/app12199548.

[5] World Bank, "Nutrition Overview: Stunting Trends." [Online]. Available: https://www.worldbank.org/en/topic/nutrition/overview. [Accessed: May 20, 2026].

[6] K. Venkatachalam, P. Trojovský, and Š. Hubálovský, "VIOLA jones algorithm with capsule graph network for deepfake detection," *PeerJ Computer Science*, vol. 9, p. e1313, 2023, doi: 10.7717/peerj-cs.1313.

[7] R. R. Reynaldo and I. Maliki, "Pengenalan Ekspresi Wajah dengan Metode Viola Jones dan Convolutional Neural Network," *Komputika*, vol. 10, no. 1, pp. 1–9, Mar. 2021, doi: 10.34010/komputika.v10i1.4119.

[8] Y. Feng, S. Yu, H. Peng, Y.-R. Li, and J. Zhang, "Detect Faces Efficiently: A Survey and Evaluations," *IEEE Trans. Biom. Behav. Identity Sci.*, vol. 4, no. 1, pp. 1–18, Jan. 2022, doi: 10.1109/TBIOM.2021.3120412.

[9] I. P. Sari, F. Ramadhani, A. Satria, and D. Apdilah, "Implementasi Pengolahan Citra Digital dalam Pengenalan Wajah menggunakan Algoritma PCA dan Viola Jones," *Hello World: Jurnal Ilmu Komputer*, vol. 2, no. 3, pp. 103–111, 2023, doi: 10.56211/helloworld.v2i3.346.

[10] Y. S. Marcelino and A. Solichin, "Deteksi Mata Katarak Berdasarkan Tekstur Gray Level Co-Occurrence Matrix Dengan Metode Self Organizing Map," *PETIR*, vol. 16, no. 2, pp. 246–256, 2023, doi: 10.33322/petir.v16i2.2104.

[11] I. D. Mienye and T. G. Swart, "A Comprehensive Review of Deep Learning : Architectures , Recent Advances , and Applications," *Information*, vol. 15, no. 12, p. 755, 2024, doi: 10.3390/info15120755.

[12] T. Ardiansyah, S. P. Balqis, and J. Haqqoni, "Deteksi Wajah dalam Foto Menggunakan Teknologi Visi Komputer," *Mars: Jurnal Teknik Informatika dan Sistem Informasi*, vol. 2, no. 6, pp. 32–39, 2024, doi: 10.61132/mars.v2i6.490.

[13] A. M. Marhelio, Munir, and Y. Wihardi, "Klasifikasi Pose Kepala Siswa Menggunakan EfficientNetV2 dengan Seat Position Embedding," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 12, no. 3, pp. 272–284, 2025.

[14] A. Ivan, T. Jaya, P. Puspitaningayu, A. P. Adiwangsa, and N. Funabiki, "Two-dimensional Human Pose Estimation using Key Points Angular Detection for Basic Strength Training," *Journal of Intelligent System and Telecommunications*, vol. 1, pp. 105–119, 2024.

[15] Y. Lu and C. Yang, "Influence of GLCM texture parameters on lithological mapping using

Sentinel-1 imagery mapping using Sentinel-1 imagery," *Geocarto International*, vol. 39, no. 1, p. 2425183, 2024, doi: 10.1080/10106049.2024.2425183.

[16] N. A. Ujilast, N. S. Firdausita, C. S. Kusumaaditya, and Y. Azhar, "MRI Image Based Alzheimer ' s Disease Classification Using," *JURNAL RESTI*, vol. 8, no. 1, pp. 18–25, 2026, doi: 10.29207/resti.v8i1.5445.

[17] P. I. ASHURI, I. A. CAHYANI, C. SRI, and K. ADITYA, "Klasifikasi Penyakit Stunting Menggunakan Algoritma Multi-Layer Perceptron," *MIND (Multimedia Artificial Intelligent Networking Database) Journal*, vol. 9, no. 1, pp. 52–63, 2024, doi: 10.26760/mindjournal.v9i1.52-63.

[18] Roboflow, "STUNTING Computer Vision Dataset." [Online]. Available: https://universe.roboflow.com/test-bdpwd/stunting-onvws. [Accessed: May 20, 2026].

[19] Roboflow, "STUNTING Computer Vision Dataset B." [Online]. Available: https://universe.roboflow.com/test-bdpwd/stunting-onvws-b12p5. [Accessed: May 20, 2026].

[20] Roboflow, "STUNTING Computer Vision Dataset C." [Online]. Available: https://universe.roboflow.com/database-ayu/deteksi-stunting. [Accessed: May 20, 2026].

[21] UNICEF Data, "Child Malnutrition – Stunting Prevalence." [Online]. Available: https://data.unicef.org/topic/nutrition/malnutrition. [Accessed: May 20, 2026].

[22] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, "A review of convolutional neural networks in computer vision," Artificial Intelligence Review, vol. 57, no. 4, p. 99, 2024, doi: 10.1007/s10462-024-10721-6.

[23] J. Ma, X. Li, Y. Ren, R. Yang, and Q. Zhao, "Landmark-Based Facial Feature Construction and Action Unit Intensity Prediction," Mathematical Problems in Engineering, vol. 2021, Article ID 6623239, pp. 1–12, 2021, doi: 10.1155/2021/6623239.

[24] J. Jiang, C. Wang, X. Liu, and C. Science, "Deep Learning-based Face Super-resolution : A Survey," ACM Computing Surveys, vol. 55, no. 1, pp. 1–36, 2021, doi: 10.1145/3485132.

[25] Q. Feng, X. Xu, and Z. Wang, "Deep learning-based small object detection : A survey," Mathematical Biosciences and Engineering, vol. 20, no. 4, pp. 6551–6590, 2023, doi: 10.3934/mbe.2023282.

[26] F. T. Kurniati, I. Sembiring, A. Setiawan, I. Setyawan, and R. R. Huizen, "GLCM-Based Feature Combination for Extraction Model Optimization in Object Detection Using Machine Learning," Jurnal Ilmiah Teknik Elektro Komputer dan Informatika, vol. 9, no. 4, pp. 1196–1205, 2023, doi: 10.26555/jiteki.v9i4.27842.

[27] G. C. Id, Y. Peng, and X. Luo, "ALGD-ORB : An improved image feature extraction algorithm with adaptive threshold and local gray difference," pp. 1–27, 2023, doi: 10.1371/journal.pone.0293111.

[28] Y. Yunidar, M. Melinda, and M. Irhamsyah, "Object Segmentation in Stunted Face Images using Deeplabv3 + with Resnet-50," *Jurnal Nasional Teknik Elektro*, vol. 13, no. 3, 2024, doi: 10.25077/jnte.v13n3.1253.2024.

[29] Y. Yunidar, Y. Yusni, N. Nasaruddin, and F. Arnia, "CNN Performance Improvement for Classifying Stunted Facial Images Using Early Stopping Approach," *JURNAL RESTI*, vol. 9, no. 1, pp. 62–68, 2026, doi: 10.29207/resti.v9i1.6068.

[30] A. Minarno, L. Wandani, and Y. Azhar, "Classification of Breast Cancer Based on Histopathological Image Using EfficientNet-B0 on Convolutional Neural Network," *International Journal of Emerging Technology and Advanced Engineering*, vol. 12, pp. 70–77, 2022, doi: 10.46338/ijetae0822_09.

[31] R. S. Das, "TensorFlow: Revolutionizing Large-Scale Machine Learning in Complex Semiconductor Design," *International Journal of Computing and Engineering*, vol. 5, no. 3, pp. 1–9, 2024.

[32] Y. Dai and J. Wu, "An Improved ORB Feature Extraction Algorithm Based on Enhanced Image and Truncated Adaptive Threshold," *IEEE Access*, vol. 11, pp. 32073–32081, 2023, doi: 10.1109/ACCESS.2023.3261665.

[33] N. P. Poetri, "Klasifikasi dan Ekstraksi Ciri Pada Jenis Kulit Wajah Dengan Metode Naïve Bayes dan Metode Gray Level Co-Occurrence Matrix (GLCM)," Bachelor thesis, Universitas Muhammadiyah Malang, Malang, Indonesia, 2023.

[34] Y. Azhar, M. R. Asyhari, V. R. S. Nastiti, A. E. Minarno, and D. R. Akbi, "Enhancing texture-based image retrieval using GLCM and DBSCAN on a multifaceted dataset," *AIP Conference Proceedings*, vol. 3179, p. 070001, 2025, doi: 10.1063/5.0259011.