Indonesian Journal of Data and Science

Volume 7 Issue 2
ISSN 2715-9930
https://doi.org/10.56705/ijodas.v7i2.528

Research Article

Leakage-Aware BISINDO Alphabet Classification with Hand-Region DINOv2 Embeddings

Lilis Nur Hayati[1]; Dolly Indra[2]

[1] Universitas Muslim Indonesia, Makassar 90231, Indonesia, lilis.nurhayati@umi.ac.id
[2] Universitas Muslim Indonesia, Makassar 90231, Indonesia, dolly.indra@umi.ac.id

Correspondence should be addressed to: lilis.nurhayati@umi.ac.id

Received 10 January 2026 Accepted 17 July 2026; Published 31 July 2026

© Authors 2026. CC BY-NC 4.0 (non-commercial use with attribution, indicate changes).
License: https://creativecommons.org/licenses/by-nc/4.0/ — Published by Indonesian Journal of Data and Science

Abstract:

Introduction: Visual similarity across images can bias the evaluation of Indonesian Sign Language (BISINDO) alphabet classifiers. This study assessed frozen DINOv2 embeddings and illumination correction using a group-disjoint evaluation protocol. **Method**: The dataset comprised 19,748 annotated gesture regions from 19,746 images across 26 classes. Full-frame visual-similarity groups were assigned to training, validation, and test partitions containing 14,107, 2,821, and 2,820 instances, respectively. Frozen DINOv2-small embeddings with class-balanced multinomial logistic regression were evaluated under raw, fixed-gamma, and adaptive-gamma conditions against a grayscale reference using principal component analysis and logistic regression. Configuration selection used validation macro-F1, while uncertainty was estimated using 2,000 group-bootstrap resamples. **Results and Discussion**: The validation-selected raw configuration achieved test accuracy of 90.28% and macro-F1 of 0.8982 (95% CI, 0.8832–0.9130), compared with macro-F1 of 0.4383 for the grayscale reference. Adaptive gamma achieved macro-F1 of 0.9038, but its advantage over raw inputs was small and uncertain (difference, 0.0056; 95% paired CI, −0.0048 to 0.0152). Exploratory brightness tests indicated condition-specific responses to synthetic illumination changes. **Conclusion**: Frozen DINOv2 embeddings supported static BISINDO gesture-region classification within an auditable evaluation framework. Adaptive gamma did not demonstrate a consistent improvement, and the findings remain limited to annotation-defined regions under the documented grouping protocol.

Keywords: BISINDO; sign-language recognition; DINOv2; hand-region classification; group-disjoint evaluation; illumination robustness.

1. Introduction

The public Multimodal BISINDO Corpus provides annotated images and videos for Indonesian Sign Language hand-gesture research [1]. Two recent BISINDO studies motivate, but do not directly benchmark against, the present work. Hayati et al. [2] evaluated gamma-corrected SSD MobileNetV2 FPNLite with 320 × 320 inputs, whereas Hayati et al. [3] formulated the task as Faster R-CNN detection. Both differ from an annotation-defined crop classifier in prediction unit, input region, and reported endpoint. Indonesian sign-language work has also used convolutional and transfer-learning approaches [4], [5], while reviews emphasize that task definition, data variation, and evaluation design determine how an accuracy claim should be interpreted [6].

Frozen visual representations provide a controlled way to separate a fixed representation from end-to-end training choices. DINOv2 learns visual features without target-task labels [7] and builds on self-supervised Vision Transformer work [8]; Vision Transformers provide the architectural basis [9], and contrastive learning supplies complementary transferable-representation evidence [10]. This study uses a frozen DINOv2-small encoder, rather than proposing a new foundation model.

Illumination correction is a focused experiment within that protocol. The fixed-gamma direction follows the preceding BISINDO illumination study [2], adaptive gamma is an image-enhancement principle [11], and photometric transformation is a recognised modelling choice [12]. Class support is unequal, so balanced training and class-sensitive reporting are motivated by imbalance literature [13], [14], macro-F1 [15], and balanced accuracy [16]. Calibration is assessed with reliability evidence and expected calibration error [17], while a group bootstrap supplies dependence-aware uncertainty [18].

The methodological gap is not simply a different network. Existing BISINDO reports do not provide a common, annotation-defined crop classification endpoint with a full-frame visual-similarity grouping audit and a locked group-disjoint partition manifest. Near duplicates can make visual-recognition evaluation optimistic [19]; perceptual hashes are practical screening tools [20], and grouped validation is appropriate when observations are structurally dependent [21]. The resulting groups control only the documented full-frame visual similarities. They do not prove independence of signers, sessions, backgrounds, or all capture conditions, and they do not demonstrate that leakage occurred in a source split. Separating development selection from the final evaluation also limits selection bias [22].

Cross-domain local studies are cited only for their limited protocol lessons: detector versus region-classifier task definitions [23], [24], controlled representation and preprocessing comparisons [25], [26], [27], and [28], and validation-leakage and multiclass reporting practice [29], [30]. No performance value is transferred across those application domains. Faster R-CNN [31] and MobileNetV2 [32] provide model-family context; class-balanced learning [33] motivates explicit support-aware reporting, while the implemented weighting rule is specified below. Two additional local studies co-authored by Huzain Azis illustrate why a frozen feature-extractor comparison [34] and imbalance-handling choices [35] should be documented explicitly; their data, models, and performance values are not transferred to the present BISINDO endpoint.

This study asks two related questions: which frozen configuration is selected by validation macro-F1, and what paired final-partition difference is observed between pre-specified raw and adaptive-gamma conditions? Its contributions are an auditable full-frame grouping procedure and locked partition; an annotation-defined BISINDO hand-region classification protocol; a scope-limited raw/fixed/adaptive gamma comparison against a simple grayscale reference; and group-bootstrap, calibration, classwise, luminance, brightness-stress, and error evidence. The supporting materials make the protocol inspectable without claiming an end-to-end recognition benchmark.

2. Materials and Methods

2.1. Corpus, unit of analysis, and partitioning

The study used version 2 of the public BISINDO corpus [1] under its CC BY 4.0 licence. The repository description reports 19,760 extracted image samples. We verified the downloaded 235c78xbmk-2.zip archive (SHA-256 982e294bece315fe0b184cb3849df62b799f4eb050fb76fea4bbfbad83e45795) as 19,753 JPEG files and 19,753 XML files, of which 19,746 JPEG-XML pairs could be matched by source identifier. Supplementary File S1 contains the archive record, source-file manifest, and annotation manifest that make this reconciliation inspectable. Seven JPEGs lacked a matching XML and seven XML files lacked a matching JPEG; the former were excluded because an input region could not be reconstructed. The resulting corpus comprises 19,748 valid XML <object> annotations from 19,746 frames. Two source frames each contain two valid, differently labelled objects.

An XML object is the source annotation’s complete alphabet-gesture region and is the analytical unit; it is not a claim that a crop contains one isolated anatomical hand. An object region may include both hands, clothing, or nearby context. Multiple valid objects in a frame are retained as separate object-level instances, but the source frame and every object it contains share one visual-context group and final partition.

Groups were built from complete source frames, never from crops. First, a SHA-256 digest of decoded pixels linked exact duplicates, irrespective of object label. The perceptual screen converted each frame to grayscale, resized it with LANCZOS to 9 × 8 pixels, and formed a 64-bit directional hash. To keep the deterministic screen tractable, same-label source candidates sharing at least one 16-bit hash band were considered; pairs then required Hamming distance ≤2, mean absolute difference ≤0.030 on 16 × 16 grayscale LANCZOS thumbnails, and global thumbnail SSIM ≥0.985. Passing pairs were joined transitively with union-find. There was no manual candidate review. Of 13,813 deterministic candidates, 6,197 passed the SSIM threshold; 961 are accepted high-similarity image pairs that cross the supplied source train/test labels (1,218 unique images). Thus, 961 denotes pairs, not groups or a count of manually adjudicated cases. The two frames containing differently labelled objects remain unified because every object from a common source frame inherits its group; exact duplicates can also link different labels. By design, nonidentical cross-label frames are outside the label-preserving perceptual candidate screen. The procedure therefore controls the documented similarity relations, not every possible cross-label visual correlation.

The supplied source split is retained only as provenance; the audit is a precautionary dependence screen and does not establish that leakage occurred. All eligible objects were allocated once by StratifiedGroupKFold with seed 20261103: one of seven outer folds was held out for test and one of six folds of the remainder for validation. This deterministic procedure produces the approximately 5:1:1 allocation in **Table 1**. No final visual-context group spans partitions.

Table 1. Corpus reconciliation and locked visual-context-group-disjoint partition.

Set or quantity	Hand instances	Source frames	Visual-context groups	Alphabet classes	Class support range
Repository-reported extracted images	-	19,760	-	-	-
Verified archive JPEG inventory	-	19,753	-	-	-
Matched XML source frames / eligible objects	19,748	19,746	16,161	26	-
Training	14,107	14,105	11,546	26	106-917
Validation	2,821	2,821	2,308	26	21-184
Final test	2,820	2,820	2,307	26	21-183
JPEGs excluded for no matched XML	0	7	-	-	-

Table 2. Scope of the closest BISINDO studies and the present annotation-defined crop protocol; their reported scores are not directly comparable.

Study	Corpus and prediction unit	Input and task	Partition / evaluation distinction
Hayati et al. [2]	BISINDO alphabet image samples; study-specific corpus version not reported	320 × 320 SSD MobileNetV2 FPNLite with gamma preprocessing	Reported 90:10 allocation; no visual-similarity group audit reported
Hayati et al. [3]	BISINDO alphabet images; detector-oriented formulation	Faster R-CNN detection rather than fixed annotation-crop classification	Detector endpoint; no common crop-classifier group-disjoint endpoint reported
Present study	Version-2 archive; one XML object as one gesture-region instance	Frozen DINOv2 embedding plus multinomial logistic-regression classification	Approximate 5:1:1 stratified visual-similarity-group-disjoint allocation; validation selection and final paired comparison separated



Figure 1. Raw training-frame example with its XML hand bounding box.

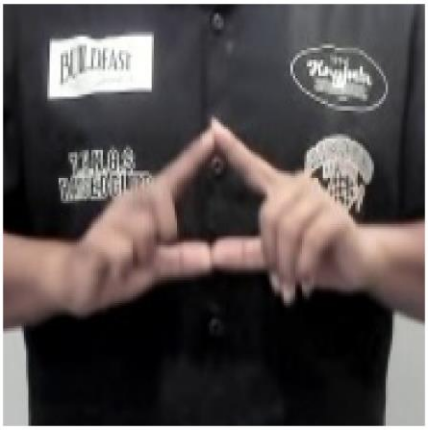


Figure 2. Raw annotation-defined hand region before geometric resizing or brightness transformation.

2.2. Exploratory data characterization

Exploratory summaries were calculated from the training partition for any quantity that could influence a modelling decision. The data contain all 26 alphabet labels but with unequal support. Bounding boxes occupy a small fraction of their source frame (median 2.25%; interquartile range 1.32%-4.75%), which motivates hand-region inputs rather than a claim about a separate full-frame benchmark. Raw crop luminance is summarised as median HSV value. Visual-context groups range from singleton observations to groups of 61 correlated instances, reinforcing the need to retain a group within one partition. Hand-region centres were concentrated near the middle of source frames (median normalised x = 0.503 and y = 0.548), while box shapes varied (median width/height ratio 0.83; interquartile range 0.68–1.02). These geometry summaries contextualise the square padded-crop procedure; neither coordinate nor aspect ratio was supplied to the classifier.

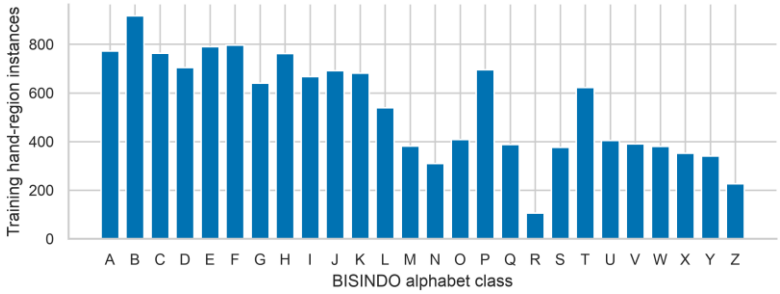


Figure 3. Training-partition class distribution for the 26 BISINDO alphabet labels.

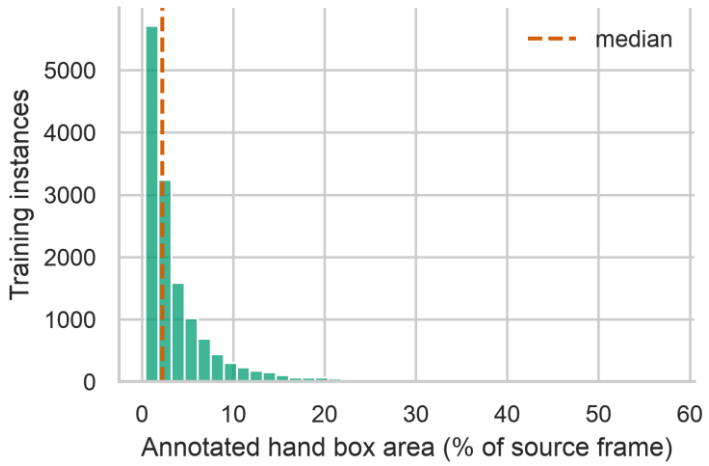


Figure 4. Training-partition distribution of annotated hand-box area as a percentage of source-frame area.

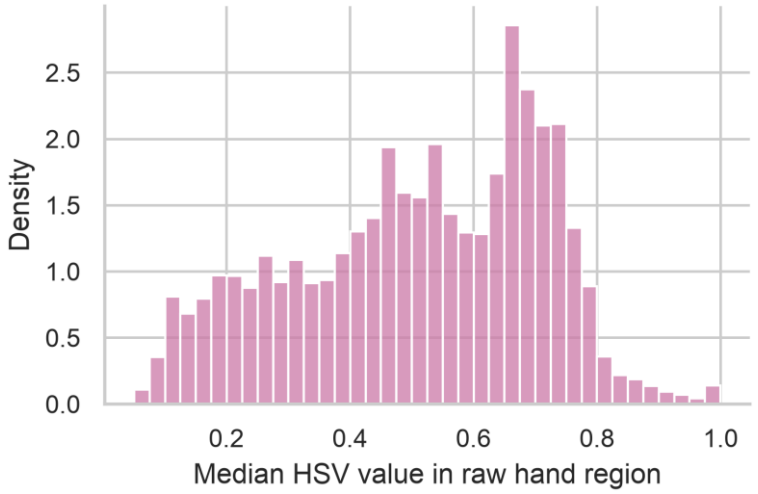


Figure 5. Training-partition distribution of raw hand-crop median HSV value.

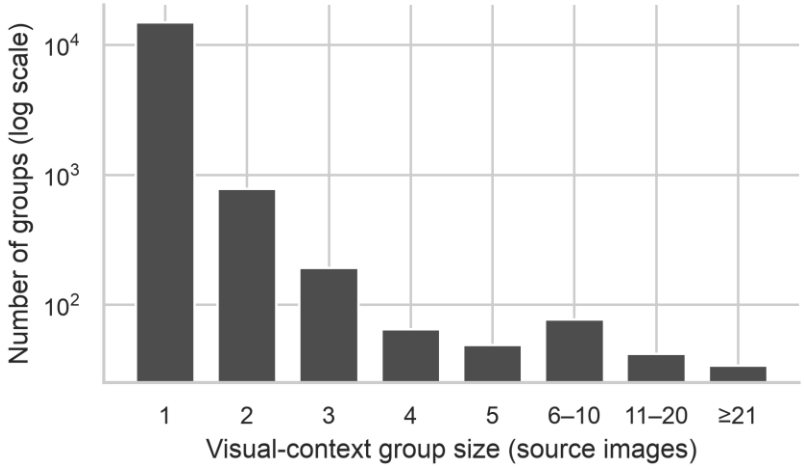


Figure 6. Distribution of visual-context group sizes after exact and perceptual similarity screening.

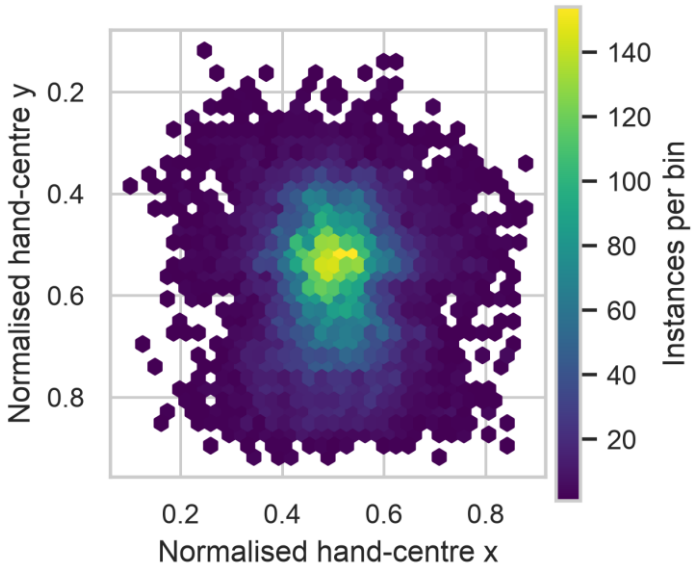


Figure 7. Spatial density of annotated hand-region centres in raw training frames; the vertical axis follows image coordinates and increases downward.

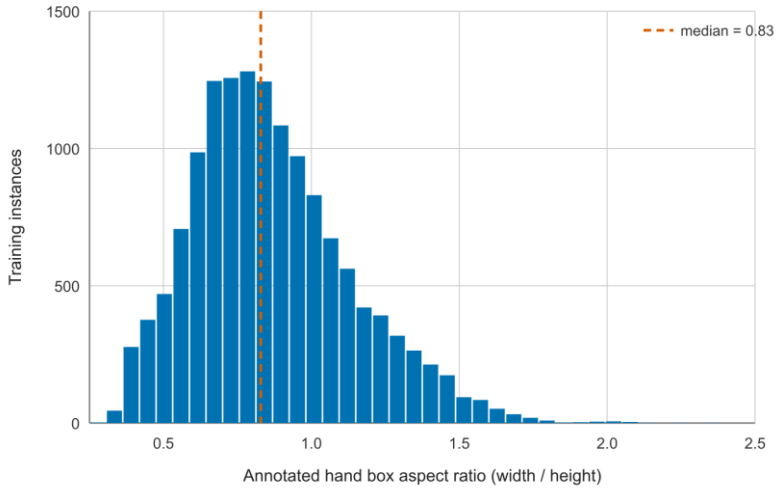


Figure 8. Training-partition distribution of annotated hand-box aspect ratio (width / height).

2.3. Region preprocessing and feature representations

Each object box was expanded symmetrically by 15% of its larger dimension and converted to a square crop. At image boundaries, the square was shifted inward and clipped to the source frame; no synthetic padding values were introduced. The earlier term “square-padded” was an imprecise description of this same implementation, not a preprocessing change; no result was recomputed for this terminology correction. The grayscale reference converted this crop to 32 × 32 grayscale pixels with BICUBIC interpolation and divided values by 255. The DINOv2 conditions used RGB crops with the exact no-register checkpoint facebook/dinov2-small (cached revision ed25f3a31f01632728cabb09d1542f84ab7b0056). Its Dinov2Model has 12 layers, 384 hidden units, six attention heads, and 14-pixel patches. The first hidden-state token (CLS, 384 dimensions) was L2-normalized before classification.

The associated BitImageProcessor converted images to RGB, resized the shortest edge to 256 pixels with BICUBIC interpolation, centre-cropped 224 × 224 pixels, rescaled values by 1/255, and normalized channels with mean (0.485, 0.456, 0.406) and standard deviation (0.229, 0.224, 0.225). All feature configurations used StandardScaler. Principal component analysis with 128 randomized-SVD components was used only for the grayscale reference. During development, scaling and PCA were fitted on training data only; after selection, they and the classifier were refitted on combined training-plus-validation data. The classifier was scikit-learn multinomial LogisticRegression with L-BFGS solver, L2 penalty, maximum 1,200 iterations, seed 20261103, and class_weight="balanced", which weights class k by $n/(26n_k)$.

For the fixed-gamma secondary comparator, the HSV value channel was transformed with $\gamma = 1.2$, fixed in advance from the prior BISINDO direction [2]; no fixed-gamma grid was selected in this experiment. The earlier description of $\gamma = 1.2$ as the "strongest fixed-gamma direction selected during validation" was a reporting error, not a procedure change: the preserved configuration and selection records show that validation varied only $C \in \{0.1, 1, 10\}$. No gamma result was recomputed. Equation (1) is an implementation-specific adaptation of the adaptive-gamma principle in [11]. For crop i, m_i is the raw median HSV value in [0,1], $t = 0.50$, and $\varepsilon = 10^{-4}$:

$$\gamma_i = clip[\log(t) / \log(\max(m_i, \varepsilon)), 0.50, 2.00], \quad V'_i = V_i \hat{} \gamma_i \tag{1}$$

The clip fixes the allowable exponent to [0.50, 2.00]. When $m_i = 1$, the logarithmic denominator is zero and the implementation deterministically assigns the upper clipped value $\gamma_i = 2.00$. Let $z_i = f\theta(x_i)$ denote the frozen embedding and w_k, b_k the fitted linear-classifier parameters. The class posterior is

$$p(y_i = k \mid z_i) = \exp(w_k^T z_i + b_k) / \Sigma_{j=1}^{26} \exp(w_j^T z_i + b_j) \tag{2}$$

Table 3. Pre-specified representations and classifier conditions.

Configuration	Region input	Representation	Classifier	Selection role
Gray PCA + LR	Raw 32 × 32 grayscale crop	Scaled PCA, 128 components	Balanced multinomial LR	Simple reference
Raw DINOv2 + LR	Raw RGB crop	Frozen DINOv2-small CLS, 384 dimensions	Same scaled balanced LR	Candidate primary
Fixed γ DINOv2 + LR	HSV value transformed with $\gamma = 1.2$	Frozen DINOv2-small CLS	Same scaled balanced LR	Fixed secondary comparator
Adaptive γ DINOv2 + LR	HSV value transformed by Equation (1)	Frozen DINOv2-small CLS	Same scaled balanced LR	Pre-specified secondary comparator

2.4. Development, final evaluation, and diagnostics

The regularisation grid $C \in \{0.1, 1, 10\}$ was evaluated using only the training and validation partitions. Macro-F1 determined the primary configuration; all configurations selected $C = 0.1$. The frozen classifiers were then fitted on combined training-plus-validation data. The final test partition was not used for feature-model selection, gamma selection, regularisation selection, threshold selection, or split revision.

For each class k, $F1_k = 2P_kR_k/(P_k + R_k)$, and the primary endpoint was

$$F1macro = (1 / 26) \Sigma_{k=1}^{26} F1_k \tag{3}$$

Secondary summaries were top-1 accuracy, balanced accuracy, macro recall, 15-bin expected calibration error, and the non-normalized multiclass Brier score

$$BS = (1 / n) \Sigma_{i=1}^{n} \Sigma_{k=1}^{26} [p(y_i = k \mid z_i) - 1(y_i = k)]^2 \tag{4}$$

For every frozen configuration, a 2,000-resample percentile group bootstrap resampled the 2,307 final-test visual-context groups with replacement and retained every object belonging to each sampled group. Macro-F1 used the fixed 26-class label set with zero_division=0; no reported resample omitted a class. The raw-versus-correction comparison used identical group draws for both predictions, yielding a paired difference interval. These post-freeze estimates do not alter the validation-selected primary model. Diagnostics comprise a row-normalised confusion matrix, complete per-class metrics and support, a reliability diagram with bin counts, training-tertile luminance strata, and exploratory synthetic-brightness stress conditions (input gamma 0.65 and 1.55). The stress order was synthetic input gamma, condition-specific correction, frozen embedding extraction, and the already fitted classifier.

Feature extraction used PyTorch 2.14.1, Transformers 5.20.2, NumPy 2.5.4, pandas 3.0.6, scikit-learn 1.9.1, and Pillow 13.0.0 on macOS arm64 with the MPS device and batch size 64. The supplied environment metadata and code identify the frozen-model and reviewer-diagnostic environments.

3. Results

3.1. Development-only selection

Table 4 separates the validation decision from final-partition summaries. Raw DINOv2 plus logistic regression had the highest validation macro-F1 (0.9042) and was designated primary before final evaluation. Adaptive gamma did not replace that decision because its validation macro-F1 was 0.9027. The grayscale reference was much lower on validation (0.4225), but it changes colour, resolution, pretraining, and representation simultaneously. It is therefore a low-performing reference condition, not evidence that isolates a DINOv2-specific benefit.

Table 4. Validation selection and final-test summaries.

Configuration	Validation macro-F1	Test macro-F1 (95% group CI)	Test accuracy	Balanced accuracy	ECE
Gray PCA + LR	0.4225	0.4383 (0.4142-0.4585)	0.4500	0.4614	0.0216
Raw DINOv2 + LR (primary)	0.9042	0.8982 (0.8832-0.9130)	0.9028	0.8981	0.0303
Fixed γ DINOv2 + LR	0.8999	0.8996 (0.8857-0.9139)	0.9035	0.8998	0.0330
Adaptive γ DINOv2 + LR	0.9027	0.9038 (0.8915-0.9162)	0.9050	0.9038	0.0321

Table 5. Paired final-test macro-F1 differences against raw DINOv2.

Frozen comparison	Difference in macro-F1	95% paired group-bootstrap interval	Interpretation
Fixed γ - raw	+0.0015	-0.0061 to 0.0091	Small difference; interval includes zero
Adaptive γ - raw	+0.0056	-0.0048 to 0.0152	Small difference; interval includes zero

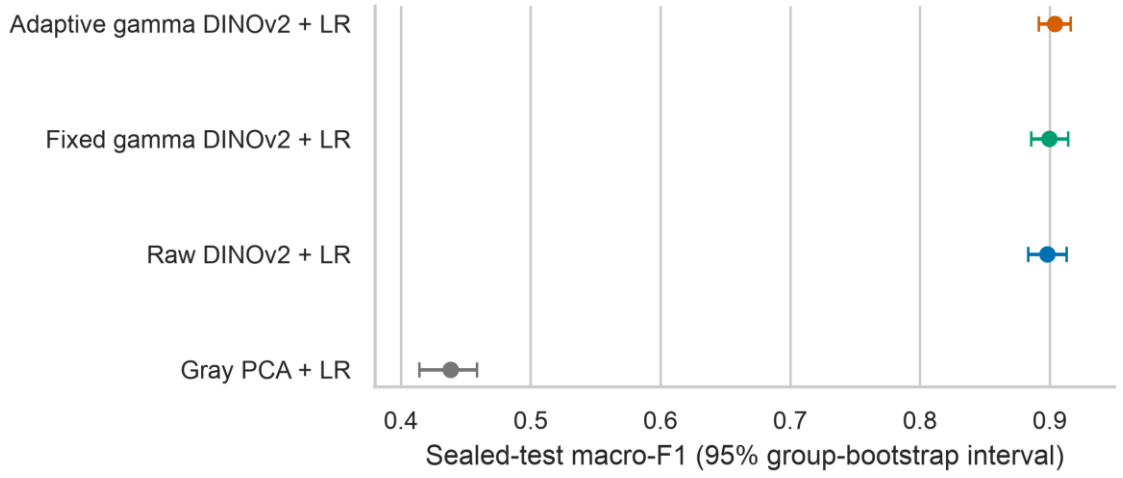


Figure 9. Final-test macro-F1 point estimates and 95% visual-context-group-bootstrap intervals for all frozen configurations.

3.2. Primary-model test performance

On the 2,820-instance final-test partition, the validation-selected raw DINOv2 primary model achieved macro-F1 0.8982 (95% group-bootstrap interval 0.8832-0.9130), balanced accuracy 0.8981, accuracy 0.9028, macro recall 0.8981, expected calibration error 0.0303, and multiclass Brier score 0.1432. The interval summarizes resampling variability within this partition, not external generalisation.

The confusion matrix shows generally high within-class recall but concentrated errors among a subset of visually similar hand shapes. The lowest classwise F1 values were P (0.7649), R (0.7692; support 21), T (0.7826), and N (0.7833); the highest were Z (1.0000; support 45), G (0.9767), F (0.9720), and Y (0.9701). Examples of the largest off-diagonal rates include N predicted as M (21.3%), R predicted as J (14.3%), M predicted as N (11.8%), and D predicted as P (10.6%).

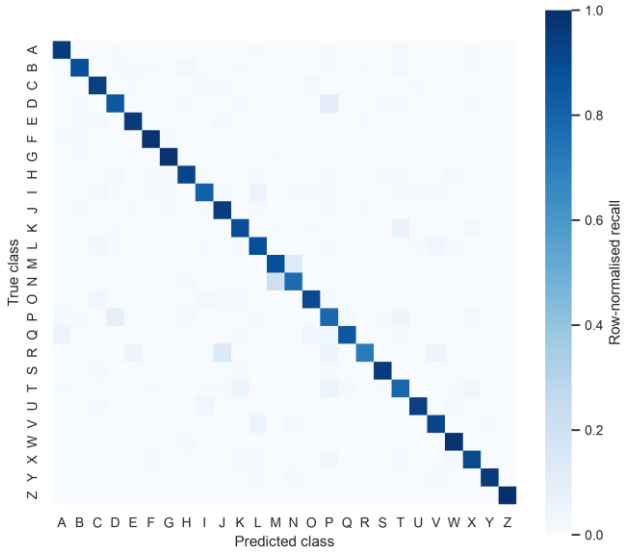


Figure 10. Row-normalised confusion matrix for the validation-selected raw DINOv2 primary classifier on the sealed group-disjoint test partition.

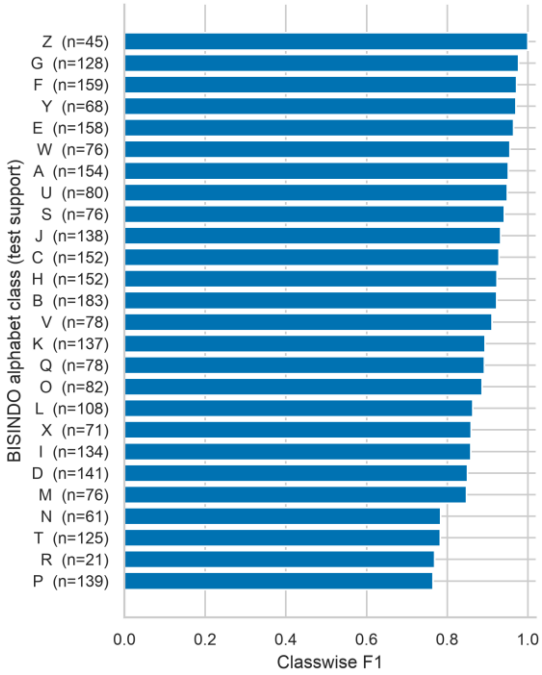


Figure 11. Per-class F1 of the validation-selected raw DINOv2 primary classifier, displayed from highest F1 at the top to lowest F1 at the bottom

Table 6. Complete final-test classwise performance for the validation-selected raw DINOv2 primary classifier.

Label	Precision	Recall	F1	Support
A	0.948	0.955	0.951	154
B	0.964	0.885	0.923	183
C	0.917	0.941	0.929	152
D	0.856	0.844	0.850	141
E	0.974	0.956	0.965	158
F	0.963	0.981	0.972	159
G	0.969	0.984	0.977	128
H	0.933	0.914	0.924	152
I	0.908	0.813	0.858	134
J	0.916	0.949	0.932	138
K	0.897	0.891	0.894	137
L	0.848	0.880	0.864	108
M	0.817	0.882	0.848	76
N	0.797	0.770	0.783	61
O	0.871	0.902	0.886	82
P	0.747	0.784	0.765	139
Q	0.943	0.846	0.892	78
R	0.833	0.714	0.769	21
S	0.924	0.961	0.942	76
T	0.773	0.792	0.783	125
U	0.962	0.938	0.949	80
V	0.900	0.923	0.911	78
W	0.926	0.987	0.955	76
X	0.821	0.901	0.859	71
Y	0.985	0.956	0.970	68
Z	1.000	1.000	1.000	45

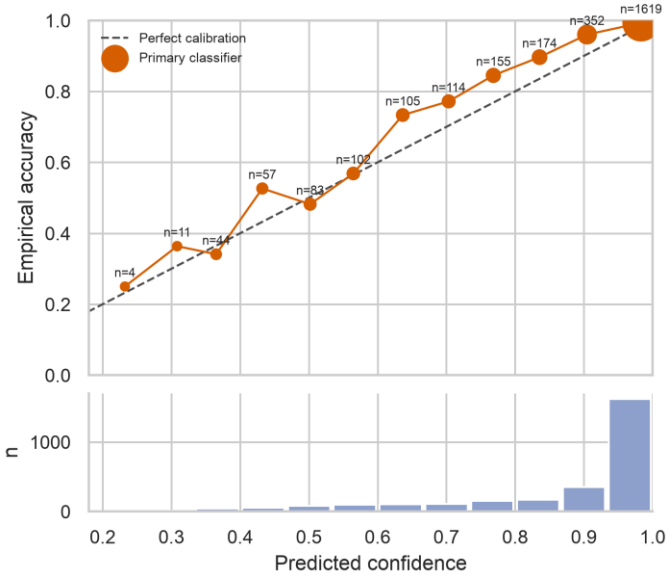


Figure 12. Reliability diagram for the validation-selected raw DINOv2 primary classifier using 15 equal-width confidence bins.

3.3. Illumination diagnostics and robustness

For the primary model, final-test macro-F1 was 0.9024 in the low-luminance stratum (n=974; median value ≤0.435294), 0.9012 in the middle stratum (n=1,011; >0.435294 to ≤0.643137), and 0.8885 in the high-luminance stratum (n=835; >0.643137). Cut points were derived from the training partition only. Supplementary Table S2 reports the complete 3 × 26 class-by-stratum count table. Different class compositions can affect stratum macro-F1, so this pattern is descriptive rather than a causal lighting estimate.

The exploratory brightness stress was extended after model freeze to all four frozen configurations. The synthetic transform was applied to the raw crop first, followed by the condition-specific correction, frozen embedding extraction, and the fitted classifier. Under brightening/darkening, respectively, macro-F1 was 0.3991/0.4433 for the grayscale reference, 0.8873/0.8816 for raw DINOv2, 0.8926/0.8787 for fixed gamma, and 0.8923/0.9011 for adaptive gamma. Thus, adaptive gamma had the largest darkened-condition score, but this post-selection diagnostic has no prospective selection role or paired uncertainty analysis.

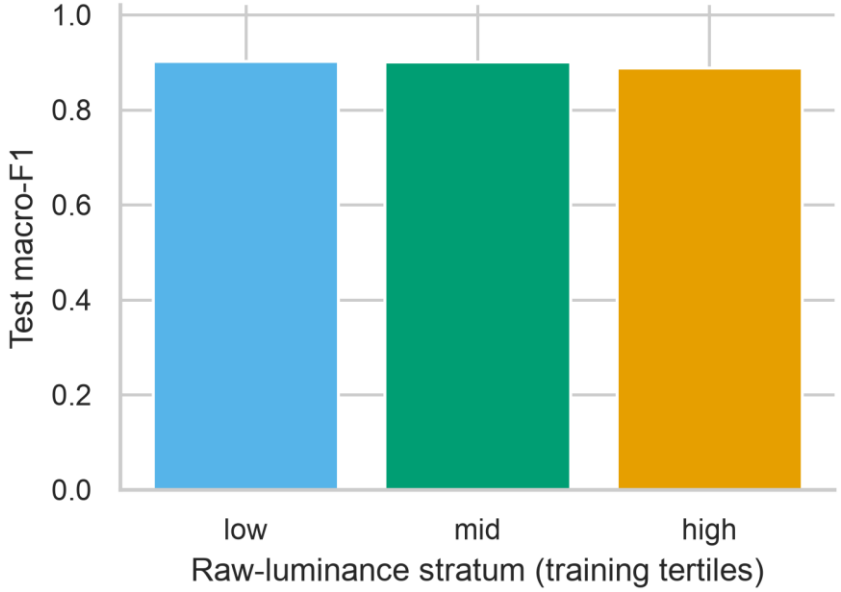


Figure 13. Primary-model macro-F1 by raw-crop luminance stratum on the final-test partition.

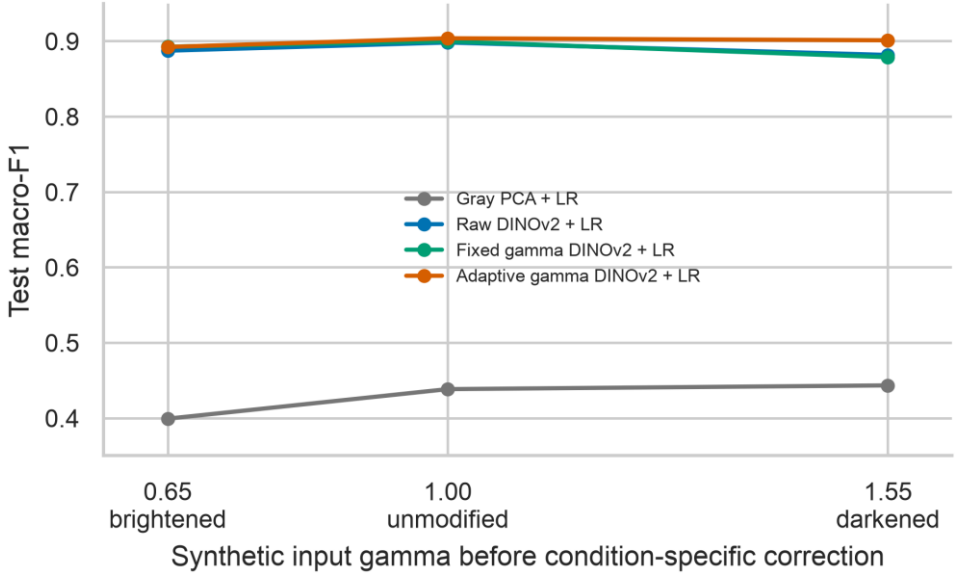


Figure 14. Exploratory post-freeze synthetic-brightness stress for every frozen condition

The large N→M and R→J rates motivated individual inspection of one high-confidence error from each pair. The annotation crops in Figures 15 and 16 are descriptive examples selected from the final-test errors; they may reflect hand-shape ambiguity, crop composition, or image quality, but they do not identify a causal error mechanism.



Figure 15. One annotation-defined final-test crop with true label N classified as M by the raw DINOv2 primary model (confidence 0.935).



Figure 16. One annotation-defined final-test crop with true label R classified as J by the raw DINOv2 primary model (confidence 0.671).

4. Discussion

The central contribution is a reconstructible evaluation protocol, not proof that leakage occurred in a prior split. The full-frame similarity screen, transitive visual-context groups, locked approximately 5:1:1 allocation, and supplied manifests reduce a documented route by which highly similar frames can cross a development/test boundary. They do not make actors, capture sessions, backgrounds, or all recording contexts independent. Within this limited estimand, the validation-selected raw DINOv2 classifier achieved macro-F1 0.8982 on the final group-disjoint partition. The study complements rather than numerically supersedes the fixed-gamma MobileNetV2 and detector-focused BISINDO studies because the unit, representation, endpoint, and partition rule differ.

The gamma comparison is scientifically modest. Adaptive gamma had the largest observed final-test macro-F1 (0.9038), 0.0056 above raw DINOv2, but raw DINOv2 retained the validation-selected primary designation. More importantly, the paired group-bootstrap interval for adaptive minus raw (-0.0048 to 0.0152) includes zero. The present experiment therefore does not establish a consistent or reliably quantified adaptive-gamma improvement. The post-freeze brightness stress suggests condition-specific behaviour under large synthetic perturbations, but it is exploratory and should be confirmed with a prospectively fixed comparison.

The grayscale PCA condition helps document a simple low-performing reference but cannot identify why DINOv2 conditions score higher: RGB input, resolution, pretraining, and representation all differ. A matched pretrained encoder with the same RGB crops, classifier, partitions, and tuning protocol remains an important future comparison. The complete classwise table, reliability-bin counts, luminance support, and individual error crops make the reported results inspectable. The two error examples should not be generalized beyond descriptive inspection.

Several limitations remain. The corpus is static and lacks suitable signer, session, and skin-tone metadata for signer-independent or fairness claims. It evaluates annotation-defined gesture regions rather than detection, continuous sign language, or communication outcomes. The visual-context grouping is deliberately conservative but cannot replace independently collected capture sessions. Finally, brightness transforms are synthetic and quantify sensitivity, not real-world illumination prevalence.

5. Conclusion

Leakage-aware BISINDO hand-region classification with frozen DINOv2 embeddings achieved macro-F1 0.8982 (95% group-bootstrap interval 0.8832-0.9130) on a 2,820-instance visual-context-group-disjoint final-test partition. The reproducible contribution delivered in Supplementary File S1 is the audit, grouping rule, manifests, implementation, and prediction outputs that define this constrained endpoint. Adaptive gamma had a small secondary advantage (+0.0056 macro-F1), but its paired group-bootstrap interval includes zero; this study does not establish a consistent improvement. Claims remain limited to static annotation-defined gesture-region classification and exclude detector, signer-independent, fairness, and continuous-sign performance.

Data Availability

The publicly licensed BISINDO corpus and its annotations are available from Mendeley Data at https://doi.org/10.17632/235c78xbmk.2 [1]. Study-specific materials are provided as Supplementary File S1: BISINDO Reproducibility Bundle with this manuscript. Its contents manifest and SHA-256 inventory enumerate the archive record; source-file, annotation, eligibility/object, grouping-audit, and locked-partition manifests; preprocessing, selection, freeze, and evaluation scripts; package and encoder metadata; validation-selection records; final per-instance predictions; numerical result tables; and figure provenance. These materials identify the source frame, XML annotation, label, group, and final partition for each eligible object. They do not redistribute the original image corpus.

References

[1] L. N. Hayati, A. N. Handayani, W. S. G. Irianto, R. A. Asmara, and D. Indra, “A Multimodal BISINDO Corpus: Annotated Images and Videos for Indonesian Sign Language Hand Gesture Recognition,” Mendeley Data, ver. 2, 2025. https://doi.org/10.17632/235c78xbmk.2.

[2] L. N. Hayati, A. N. Handayani, W. S. G. Irianto, R. A. Asmara, D. Indra, and N. S. Damanhuri, “Improving Indonesian Sign Alphabet Recognition for Assistive Learning Robots Using Gamma-Corrected MobileNetV2,” Buletin Ilmiah Sarjana Teknik Elektro, vol. 7, no. 3, pp. 350-361, 2025. https://doi.org/10.12928/biste.v7i3.13300.

[3] L. N. Hayati, A. N. Handayani, W. S. G. Irianto, R. A. Asmara, D. Indra, and N. S. Damanhuri, “Real-Time BISINDO Alphabet Recognition via Faster R-CNN Incorporating Skin Tone Diversity as a Classification Feature,” Buletin Ilmiah Sarjana Teknik Elektro, vol. 8, no. 3, pp. 811-823, 2026. https://doi.org/10.12928/biste.v8i3.15587.

[4] S. Dwijayanti, Hermawati, S. I. Taqiyyah, H. Hikmarika, and B. Y. Suprapto, “Indonesia Sign Language Recognition Using Convolutional Neural Network,” International Journal of Advanced Computer Science and Applications, vol. 12, no. 10, pp. 415-422, 2021. https://doi.org/10.14569/IJACSA.2021.0121046.

[5] S. Suharjito, N. Thiracitta, and H. Gunawan, “SIBI Sign Language Recognition Using Convolutional Neural Network Combined with Transfer Learning and Non-Trainable Parameters,” Procedia Computer Science, vol. 179, pp. 72-80, 2021. https://doi.org/10.1016/j.procs.2020.12.011.

[6] I. A. Adeyanju, O. O. Bello, and M. A. Adegboye, “Machine Learning Methods for Sign Language Recognition: A Critical Review and Analysis,” Intelligent Systems with Applications, vol. 12, Art. no. 200056, 2021. https://doi.org/10.1016/j.iswa.2021.200056.

[7] M. Oquab et al., "DINOv2: Learning Robust Visual Features Without Supervision," Transactions on Machine Learning Research, 2024. [Online]. Available: https://openreview.net/forum?id=a68SUt6zFt.

[8] M. Caron et al., "Emerging Properties in Self-Supervised Vision Transformers," in Proc. IEEE/CVF International Conference on Computer Vision, 2021, pp. 9650-9660. https://doi.org/10.1109/ICCV48922.2021.00968.

[9] A. Dosovitskiy et al., "An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale," in Proc. International Conference on Learning Representations, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy.

[10] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in Proc. 37th International Conference on Machine Learning, 2020, pp. 1597-1607. [Online]. Available: https://proceedings.mlr.press/v119/chen20j.html.

[11] S. Rahman, M. M. Rahman, M. Abdullah-Al-Wadud, G. D. Al-Quaderi, and M. Shoyaib, "An Adaptive Gamma Correction for Image Enhancement," EURASIP Journal on Image and Video Processing, vol. 2016, Art. no. 35, 2016. https://doi.org/10.1186/s13640-016-0138-1.

[12] C. Shorten and T. M. Khoshgoftaar, "A Survey on Image Data Augmentation for Deep Learning," Journal of Big Data, vol. 6, Art. no. 60, 2019. https://doi.org/10.1186/s40537-019-0197-0.

[13] J. M. Johnson and T. M. Khoshgoftaar, "Survey on Deep Learning with Class Imbalance," Journal of Big Data, vol. 6, Art. no. 27, 2019. https://doi.org/10.1186/s40537-019-0192-5.

[14] M. Buda, A. Maki, and M. A. Mazurowski, "A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks," Neural Networks, vol. 106, pp. 249-259, 2018. https://doi.org/10.1016/j.neunet.2018.07.011.

[15] M. Sokolova and G. Lapalme, "A Systematic Analysis of Performance Measures for Classification Tasks," Information Processing & Management, vol. 45, no. 4, pp. 427-437, 2009. https://doi.org/10.1016/j.ipm.2009.03.002.

[16] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The Balanced Accuracy and Its Posterior Distribution," in Proc. 20th International Conference on Pattern Recognition, 2010, pp. 3121-3124. https://doi.org/10.1109/ICPR.2010.764.

[17] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in Proc. 34th International Conference on Machine Learning, 2017, pp. 1321-1330. [Online]. Available: https://proceedings.mlr.press/v70/guo17a.html.

[18] B. Efron and R. Tibshirani, "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy," Statistical Science, vol. 1, no. 1, pp. 54-75, 1986. https://doi.org/10.1214/ss/1177013815.

[19] B. Barz and J. Denzler, "Do We Train on Test Data? Purging CIFAR of Near-Duplicates," Journal of Imaging, vol. 6, no. 6, Art. no. 41, 2020. https://doi.org/10.3390/jimaging6060041.

[20] C. Zauner, "Implementation and Benchmarking of Perceptual Image Hash Functions," M.S. thesis, Upper Austria University of Applied Sciences, Hagenberg, Austria, 2010. [Online]. Available: https://zauner.nllk.net/files/thesis.pdf.

[21] D. R. Roberts et al., "Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure," Ecography, vol. 40, no. 8, pp. 913-929, 2017. https://doi.org/10.1111/ecog.02881.

[22] S. Varma and R. Simon, "Bias in Error Estimation When Using Cross-Validation for Model Selection," BMC Bioinformatics, vol. 7, Art. no. 91, 2006. https://doi.org/10.1186/1471-2105-7-91.

[23] M. R. Romadlon and A. N. Handayani, "YOLOv8 Implementation on British Sign Language System with Edge Detection Extraction," Indonesian Journal of Data and Science, vol. 6, no. 2, pp. 211-221, 2025. https://doi.org/10.56705/ijodas.v6i2.276.

[24] M. E. Wijaya and A. N. Handayani, "Integration of YOLOv8 and Instance Segmentation in the Chinese Sign Language Recognition System," Indonesian Journal of Data and Science, vol. 6, no. 2, pp. 241-250, 2025. https://doi.org/10.56705/ijodas.v6i2.247.

[25] A. K. F. Shobiry and R. Puspitasari, "Hybrid Feature Benchmark for Blood Cell Classification Using ResNet50 and EfficientNetV2 Features with SVM and ANN Classifiers via Unsupervised Segmentation," International Journal of Artificial Intelligence in Medical Issues, vol. 3, no. 2, pp. 110-126, 2025. [Online]. Available: https://jurnal.yoctobrain.org/index.php/ijaimi/article/view/364.

[26] H. Oumarou, "Detection and Classification of Bacterial Skin Infections Using K-Nearest Neighbors Algorithm," International Journal of Artificial Intelligence in Medical Issues, vol. 1, no. 2, pp. 124-136, 2023. https://doi.org/10.56705/ijaimi.v1i2.149.

[27] A. R. Manga, T. Amaliah, Y. Salim, T. Saitoh, H. Zhang, and H. Azis, "From Embeddings to Ensembles: A Controlled Comparison of ArcFace and Global CNN Features," in 2026 20th International Conference on Ubiquitous Information Management and Communication, 2026, pp. 1-8. https://doi.org/10.1109/IMCOM69009.2026.11360880.

[28] A. R. Manga, Nirmala, H. Azis, F. Fattah, Y. Salim, and H. Darwis, "ResNet-50 for Flower Image Classification: A Comparative Study of Segmentation and Non-Segmentation Approaches," in 2025 19th International Conference on Ubiquitous Information Management and Communication, 2025, pp. 1-6. https://doi.org/10.1109/IMCOM64595.2025.10857520.

[29] N. Rismayanti, "Predicting the Rule, Not the Load: A Construct-Validity and Validation-Leakage Audit of Steel Industry Load-Type Classification," Nusantara Journal of Multidisciplinary Informatics, vol. 1, no. 1, pp. 34-41, 2026. [Online]. Available: https://njomi.org/index.php/njomi/article/view/5.

[30] N. Rismayanti and A. P. Utami, "Improving Multi-Class Classification on 5-Celebrity-Faces Dataset using Ensemble Classification Methods," Indonesian Journal of Data and Science, vol. 4, no. 2, pp. 120-129, 2023. https://doi.org/10.56705/ijodas.v4i2.78.

[31] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 2017. https://doi.org/10.1109/TPAMI.2016.2577031.

[32] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510-4520. https://doi.org/10.1109/CVPR.2018.00474.

[33] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-Balanced Loss Based on Effective Number of Samples," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9268-9277. https://doi.org/10.1109/CVPR.2019.00949.

[34] H. Azis, R. A. Jalil, and A. R. Manga, "Comparative Performance of VGG16 and EfficientNetB0-Based Transfer Learning for Brain Tumor Classification," Knowledge Engineering and Data Science, vol. 8, no. 2, pp. 185-197, 2025. https://doi.org/10.17977/um018v8i22025p185-197.

[35] Y. Salim, A. P. Utami, A. R. Manga, H. Azis, and F. T. Admojo, "Optimal Strategy for Handling Unbalanced Medical Datasets: Performance Evaluation of K-NN Algorithm Using Sampling Techniques," Knowledge Engineering and Data Science, vol. 7, no. 2, pp. 176-186, 2024. https://doi.org/10.17977/um018v7i22024p176-186.