Vision Transformer-Based Dog Breed Classification with a Hybrid Detection-Classification Framework

Njoto Benarkah[1*], Joko Siswantoro[2], Bryan Porayouw[3]

[1,2,3]Informatics Engineering, Faculty of Engineering, University of Surabaya, Surabaya, East Java, Indonesia
E-mail: [1*]benarkah@staff.ubaya.ac.id, [2]joko_siswantoro@staff.ubaya.ac.id, [3]s160422045@student.ubaya.ac.id

(Received: 30 Apr 2026, revised: 23 May 2026, accepted: 25 May 2026)

Abstract

Dog breed classification remains a challenging task in computer vision due to high inter-class visual similarity, pose variations, changes in illumination, and complex background conditions. Conventional convolutional neural network (CNN) approaches often struggle to capture global contextual dependencies and subtle discriminative features. This study proposes a hybrid deep learning framework that integrates YOLOv8n for object detection with the Vision Transformer (ViT-B/16) for dog breed classification. The dataset comprises 14,181 dog images collected from the Tsinghua Dogs Dataset and supplementary real-world sources, spanning 10 dog breed categories. The proposed framework includes image preprocessing, data augmentation, transfer learning, and Bayesian hyperparameter optimization using Optuna to enhance model generalization. YOLOv8n is employed to localize dog regions, which are subsequently resized and passed to the Vision Transformer for global feature representation learning. The model is evaluated on 2,133 unseen test images. Experimental results demonstrate that the proposed framework achieves an accuracy of 97.98% with macro and weighted F1-score values of 98.76% and 97.98%, respectively. Comparative experiments against standalone ViT-B/16 and EfficientNetV2M architectures futher confirm the effectiveness of the proposed hybrid YOLOv8n–ViT-B/16 framework for dog breed classification.

Keywords: Vision Transformer, Image Classification, Dog Breed Classification, Hybrid Deep Learning, Object Detection.

I. INTRODUCTION

Dogs are among the most widely kept companion animals worldwide [1], [2]. Each dog breed exhibits distinct morphological characteristics, including body structure, coat color, behavioral traits, and functional capabilities. These variations are associated with breed-specific health conditions [3], dietary requirements [4], and grooming needs [5]. Accurate breed identification supports veterinary decision-making by enabling personalized treatment, preventive care strategies, and clinical management of breed-associated diseases [6], [7]. Breed-level information also contributes to behavioral training, working dog selection, and animal adoption processes, where correct identification improves matching accuracy between animals and handlers [8]. Ethical considerations arise in automated recognition systems due to potential misclassification risks and the reinforcement of unintended breed-related biases, emphasizing the necessity for reliable and responsible computational approaches [6], [9].

Dog-breed recognition is considered a challenging fine-grained visual classification task due to high inter-class similarity and subtle intra-class variation. Many breeds share overlapping visual attributes, such as coat texture, body proportions, and facial structure, leading to frequent ambiguity during manual identification. Real-world imaging conditions introduce additional complexity through occlusion, illumination variation, pose changes, and background clutter, which further reduce the visibility of discriminative features. Strong visual resemblance between breeds such as Golden Retriever and Labrador Retriever exemplifies the difficulty of distinguishing closely related categories.

Large-scale breed diversity further increases complexity, with more than 180 recognized breeds exhibiting subtle but meaningful visual differences [10], [11]. Variability within the same breed adds another layer of difficulty, particularly when environmental conditions alter observable characteristics. These combined factors position dog breed recognition as a highly complex, fine-grained classification problem that requires robust feature representation and strong generalization capabilities.

Traditional Convolutional Neural Network (CNN)-based architectures have demonstrated notable improvements in dog breed classification performance. CNN-based models have achieved an accuracy of up to 96.75%, compared to 79.25% with HOG descriptors, highlighting the superiority of deep feature learning for fine-grained recognition tasks [12].

DOI: 10.34148/teknika.v15i2.1484

EfficientNet-based models further achieve approximately 90% test accuracy on multi-breed datasets through compound scaling and transfer learning strategies, demonstrating strong classification capability while still encountering limitations in modeling subtle inter-class variations [13]. Earlier architectures, such as InceptionV3, reported approximately 78% accuracy on the Stanford Dogs dataset, revealing challenges related to overfitting and limited generalization in conventional CNN pipelines [8].

Transformer-based architectures have recently emerged as powerful alternatives in computer vision due to their ability to model long-range dependencies through self-attention mechanisms [14], [15]. Vision Transformer (ViT) reformulates image recognition as a sequence modeling problem by dividing images into patches and learning global relationships across them [16]. Global contextual modeling enables more effective discrimination of subtle visual differences than conventional localized convolutional operations. ViT is generally more robust to image perturbations than CNNs, making it suitable for safety-critical applications [17], [18].

Recent studies demonstrate strong performance of transformer-based models in fine-grained recognition tasks. Canto et al. [19] demonstrated that Vision Transformer-based approaches outperform ResNet-based CNN architectures in dog-face recognition, confirming the effectiveness of self-attention mechanisms in capturing spatial dependencies for identity-level classification tasks. Multimodal learning frameworks, such as CLIP-ASN, integrate visual and auditory information using transformer-based contrastive learning and co-attention mechanisms, achieving 89.75% accuracy and improving robustness under degraded input conditions [9]. Lee et al. proposed a bi-directional attention framework for zero-shot learning, where transformer-inspired mechanisms enhance visual-semantic alignment and improve discrimination among closely related categories, achieving strong results on fine-grained benchmarks such as CUB.

Recent findings indicate that attention-based transformer architectures offer strong capabilities for modeling fine-grained visual distinctions and improving class separability. This property supports the suitability of transformer-based models for dog-breed recognition, where subtle inter-class differences are critical for accurate classification.

Previous studies have demonstrated the effectiveness of CNN-based and transformer-based architectures for dog breed recognition. Several limitations remain insufficiently addressed. Conventional CNN approaches primarily focus on localized feature extraction and often struggle to model long-range contextual dependencies among visually similar breeds. Transformer-based methods improve global feature representation, yet most existing studies rely on classification architectures without explicitly isolating dog regions from complex background before classification. Existing approaches also provide limited integration between object localization and transformer-based semantic representation learning for dog breed recognition.

This study proposes a hybrid framework for fine-grained dog breed classification that integrates YOLOv8n object detection [20] with a Vision Transformer (ViT-B/16) classifier. YOLOv8n is utilized to localize dog regions and suppress background noise, while the Vision Transformer performs global feature learning for classification. The proposed integration enhances feature robustness and improves classification performance across visually similar dog-breed categories through the combination of localized object detection and global contextual representation learning.

II. RESEARCH METHODOLOGY

A. Dataset

This study used a dataset of 14,181 dog images collected from the Tsinghua Dog Dataset [21], supplemented with additional images from Google Images and direct photography. The dataset covered 10 dog breeds: Golden Retriever, Labrador Retriever, German Shepherd, Rottweiler, Boxer, Siberian Husky, Samoyed, Bernese Mountain Dog, Great Dane, and Doberman.

Table 1. Dataset Distribution of Dog Breed Images

Dog Breed	Tsinghua	Google	Photo	Total
Golden Retriever	5,355	-	-	5,355
Labrador Retriever	3,576	-	4	3,580
German Shepherd	208	75	7	290
Rottweiler	224	76	-	300
Boxer	225	101	-	326
Siberian Husky	1,160	-	-	1,160
Samoyed	2,177	-	15	2,192
Bernese Mt Dog	211	112	-	323
Great Dane	206	160	-	366
Doberman	209	80	-	289
Total	13,551	604	26	14,181

Table 1 presents the distribution of dog breed images across different data sources. The Tsinghua Dog Dataset provided the majority of samples, particularly for high-frequency classes such as Golden Retriever, Labrador Retriever, Siberian Husky, and Samoyed. Supplementary images from Google Images and direct photography were incorporated to enhance class balance and increase visual diversity in underrepresented breeds, including German Shepherd, Rottweiler, Boxer, Great Dane, and Doberman. This combination improves dataset variability and strengthens model generalization under real-world conditions.

The dataset selection represents a controlled subset of the full Tsinghua Dogs Dataset, focusing on ten breeds with strong inter-class visual similarity to formulate a challenging fine-grained classification task. A veterinary expert verified all annotations to ensure label consistency and breed correctness. The dataset was partitioned into training, validation, and testing subsets using a stratified split ratio of 70%, 15%, and 15%, respectively.

DOI: 10.34148/teknika.v15i2.1484

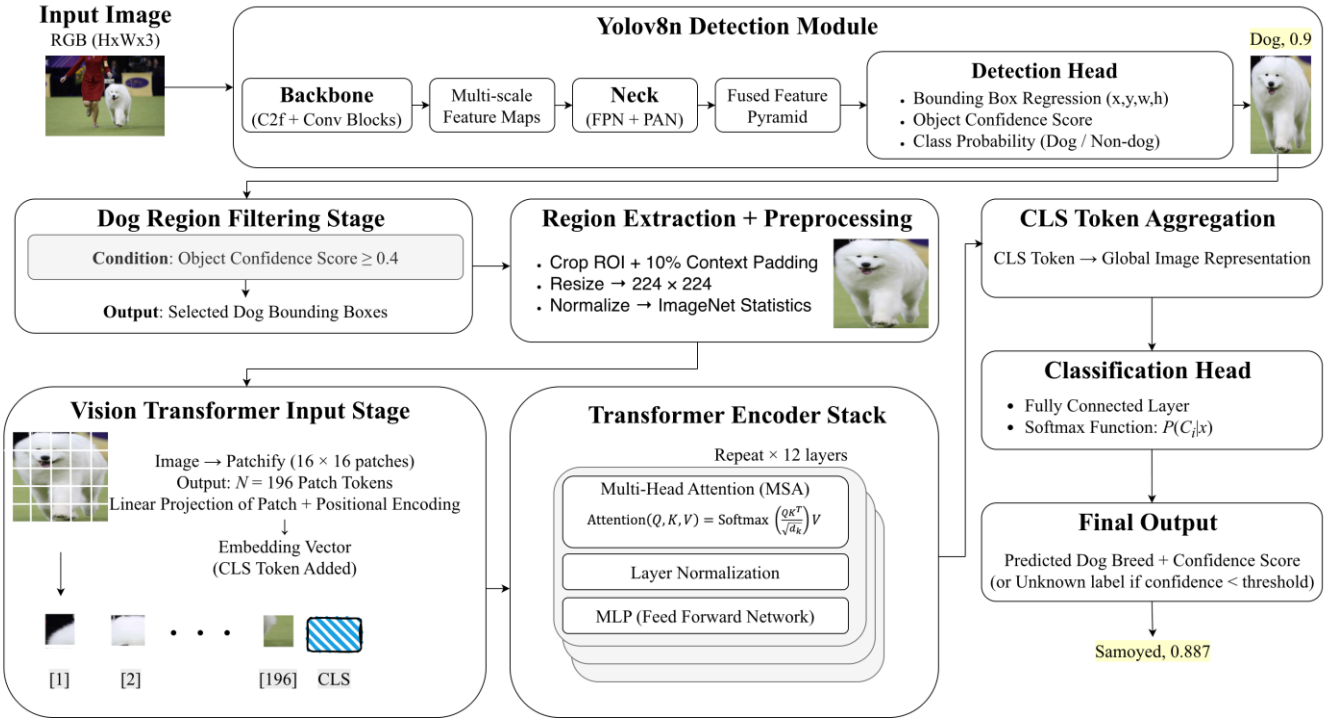


Figure 1. Proposed Hybrid YOLOv8n–ViT-B/16 Framework for Dog Breed Classification

B. Proposed Framework

This study proposes a hybrid vision framework that integrates object detection and Transformer-based classification for dog breed recognition. The proposed pipeline consists of five sequential stages: object detection using YOLOv8n, region extraction and preprocessing, Vision Transformer-based feature learning, hyperparameter optimization, and performance evaluation. Figure 1 illustrates the complete workflow of the proposed hybrid YOLOv8n–ViT-B/16 framework for dog breed classification. YOLOv8n performs spatial localization of dog objects, while the Vision Transformer (ViT-B/16) performs breed-level classification using global contextual representations.

The YOLOv8n module processes the input RGB image through a hierarchical convolutional backbone composed of C2f and convolutional layers to extract multi-scale feature representations. These features are aggregated using a neck structure based on Feature Pyramid Network (FPN) [22], [23] and Path Aggregation Network (PAN) [24]. The detection head generates bounding boxes, object confidence scores, and class probabilities. Regions satisfying the confidence threshold of 0.40 are retained for subsequent processing. This filtering stage reduces irrelevant background regions and improves the quality of extracted object features.

The Vision Transformer (ViT-B/16) module receives cropped dog regions resized to 224 × 224 pixels. Each image is partitioned into fixed-size patches of 16 × 16 pixels, resulting in a sequence of $N = 196$ patch tokens. Each patch is flattened and linearly projected to generate latent embedding vectors. Positional encoding preserves spatial relationships among patches, while the classification token (CLS token) aggregates global semantic information. The patch embedding process is formulated in Equation 1.

$$z_0 = \left[x_{cls}; x_p^1 E; x_p^2 E; \ldots; x_p^N E\right] + E_{pos} \quad (1)$$

Equation 1 defines the patch embedding sequence z_0 used as the input of the Transformer encoder. Variable z_0 represents the combined patch embedding sequence consisting of image patch embeddings, positional embeddings, and the classification token. Variable x_{cls} represents the classification token employed for global feature aggregation. Variable x_p^i denotes the i-th patch extracted from the input image. Matrix E represents the learnable linear projection matrix that transforms image patches into embedding vectors. Variable E_{pos} denotes the positional embedding that preserves spatial relationships among patch tokens.

The embedded sequence is processed through multiple Transformer encoder layers, each consisting of multi-head self-attention (MSA) [25] and multilayer perceptron (MLP) blocks.

The final output of the encoder is represented by the classification token (CLS), which aggregates global semantic information across all patches. This representation is forwarded to a fully connected layer followed by a Softmax classifier to produce probability distributions over ten dog breed categories.

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

Equation 2 defines the self-attention mechanism employed in the Vision Transformer architecture. The self-attention

DOI: 10.34148/teknika.v15i2.1484

mechanism computes relationships between image patches using query (Q), key (K), and value (V) matrices. These matrices are linear transformations of the input embedding, where Q encodes the feature representations used to query relevant information between image patches. Key matrix K encodes reference feature descriptors used to compute similarity scores. Value matrix V contains the corresponding feature representations that contribute to the final attention output. The term QK^T computes similarity scores between query and key matrices, while the scaling factor $\sqrt{d_k}$ stabilizes gradient variance during training. The Softmax function normalizes the similarity scores into attention weights, enabling aggregation of global contextual information across regions. The normalized similarity scores determine the contribution of each image patch to the final feature representation.

The proposed framework adopts a sequential hybrid design that combines YOLOv8n for spatial localization and Vision Transformer for semantic classification. The YOLOv8n detector performs object localization by filtering out irrelevant background regions and extracting only dog-specific regions. These regions are then normalized and passed to the ViT module for global feature modeling.

The Vision Transformer enhances classification performance by capturing long-range dependencies among image patches through self-attention mechanisms. This enables the model to distinguish subtle inter-class variations among visually similar dog breeds. The final prediction is obtained from the CLS token representation, which is processed by a fully connected layer with Softmax activation to produce the predicted breed label and a confidence score.

The hybrid integration of YOLOv8n and ViT-B/16 provides a complementary learning framework that leverages both localized object detection and global feature modeling. This integrated design improves robustness in handling intra-class similarity and inter-class variability.

C. Hyperparameter Optimization

This study employed hyperparameter optimization to determine the optimal configuration for the Vision Transformer (ViT-B/16) classification model. The optimization process utilized the Optuna framework with the Tree-structured Parzen Estimator (TPE) Bayesian Optimization to efficiently explore the hyperparameter search space [26], [27]. The optimization workflow consisted of sequential stages, including search space definition, hyperparameter generation, model training, validation evaluation, early stopping and pruning, and best parameter selection.

The optimization process explored several hyperparameters affecting the training dynamics and generalization capability of the Vision Transformer model. The search space included the learning rate, weight decay, batch size, label smoothing factor, layer-freeze ratio, and optimizer type. Learning rate values were sampled logarithmically within the range of 1×10^{-5} to 8×10^{-5}, while weight decay values ranged from 0.01 to 0.08. Batch size selection consisted of 16 and 32 samples per iteration. Label smoothing values varied between 0.05 and 0.15 to reduce model overconfidence during classification. Freeze ratio values ranged from 0.5 to 0.75 to control the proportion of frozen Transformer layers during fine-tuning. The optimizer configuration evaluated both Adam and AdamW optimization algorithms.

The optimization procedure employed TPE Bayesian Optimization to generate adaptive hyperparameter combinations based on the performance of previous trials. Each trial initialized the ViT-B/16 model with a sampled hyperparameter configuration and then performed training and validation. The training stage utilized cross-entropy loss with label smoothing, gradient clipping, and cosine annealing learning rate scheduling with warmup. Equation 3 defines the cross-entropy loss with label smoothing employed during training.

$$\mathcal{L} = -\sum_{i=1}^{C} y_i^{LS} \log(p_i) \tag{3}$$

Variable C represents the total number of classes, p_i denotes the predicted probability for class i, and y_i^{LS} represents the smoothed ground-truth label distribution. The label smoothing process reduces overconfidence by distributing a small probability mass to non-target classes.

The learning rate scheduling mechanism used cosine annealing with a warmup to stabilize optimization during the early training epochs. Equation 4 defines the cosine learning rate schedule.

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min})\left(1 + \cos\left(\frac{t}{T}\pi\right)\right) \tag{4}$$

Variable η_t denotes the learning rate at iteration t, T represents the total training iterations, while η_{max} and η_{min} denote the maximum and minimum learning rates, respectively.

The optimization process incorporated early stopping and Optuna pruning mechanisms to reduce computational overhead and prevent overfitting. Early stopping terminated training when validation loss failed to improve for five consecutive epochs. Optuna pruning discontinued underperforming trials after several training epochs based on the progression of validation loss. Each trial recorded training loss, validation loss, training accuracy, validation accuracy, learning rate progression, and generalization gap values for performance analysis. The final hyperparameter configuration was selected according to the highest validation accuracy achieved across 25 optimization trials.

D. Model Training

The training process utilized transfer learning with a pretrained Vision Transformer (ViT-B/16) model. Data augmentation techniques were applied during preprocessing to improve model generalization, including random cropping, horizontal flipping, rotation, color jittering, affine transformations, and random erasing. Class imbalance was partially mitigated through augmentation strategies and stratified dataset partitioning, which preserved class

DOI: 10.34148/teknika.v15i2.1484

distribution consistency across training, validation, and testing subsets. The augmentation operations increased visual diversity and reduced model overfitting, particularly for minority classes with limited training samples.

All input images were resized to 224×224 pixels and normalized using ImageNet mean and standard deviation statistics to ensure consistency with the pretrained ViT backbone. The forward propagation stage generated class-probability outputs for dog-breed classification, while the backward propagation stage updated model parameters using the AdamW optimizer.

The optimization process incorporated a cosine-annealing learning rate scheduler with 10% warmup steps, enabling stable convergence during the early training epochs. The loss function was defined using cross-entropy loss with label smoothing, which reduced overconfidence and improved generalization across visually similar dog breeds.

Model training was conducted using a mini-batch strategy, and performance was evaluated after each epoch using validation loss and validation accuracy. The training process employed gradient clipping with a max-norm of 1.0 to stabilize updates and prevent gradient explosion.

An early stopping mechanism with a patience of five epochs was implemented to terminate training when validation loss failed to improve. The model checkpointing strategy stored the best-performing model based on minimum validation loss, ensuring optimal generalization performance.

E. Dog Breed Classification Process

The classification process began when the system received an input image. The YOLOv8n object detection model identified dogs in the image using pretrained detection weights. Bounding boxes with confidence scores greater than or equal to 0.4 were selected for further processing.

The selected regions undergo cropping with 10% contextual padding to preserve full object boundaries and avoid feature loss. The cropped regions are resized to 224 × 224 pixels and normalized using ImageNet statistics before being passed to the classification stage. The Vision Transformer (ViT-B/16) module processes the cropped regions by dividing each image into fixed-size patches of 16 × 16 pixels, forming a sequence of patch tokens. These tokens are embedded and processed through transformer encoder layers to capture global contextual relationships.

The classification head generated probability distributions over all dog breed categories. The final prediction is determined based on confidence scores greater than or equal to 0.85; lower-confidence predictions receive an unknown label.

F. Model Evaluation

The evaluation process utilized 2,133 test images that were strictly excluded from the training and validation stages to ensure unbiased performance assessment.

Model performance was evaluated using multiple metrics, including accuracy, precision, recall, F1-score, classification report, and confusion matrix. These metrics provide a comprehensive evaluation of both overall classification performance and class-wise prediction behavior.

The confusion matrix was used to analyze inter-class misclassification patterns, particularly among visually similar dog breeds. This enabled detailed analysis of class confusion patterns and model discrimination capability.

Training dynamics were further evaluated using learning curves for training and validation loss and training and validation accuracy, which were recorded at each epoch. These curves were used to assess convergence behavior, the generalization gap, and the tendency toward overfitting throughout the training process.

Fair experimental comparison was maintained by training and testing all evaluated models using identical training, validation, and testing partitions generated through the same stratified split strategy. Comparative baseline models, including standalone ViT-B/16 and EfficientNetV2M, utilized the original uncropped images, while the proposed framework utilized YOLOv8n-cropped dog regions prior to classification. The experimental design enables direct evaluation of the contribution of the YOLOv8n localization stage to overall classification performance.

G. Implementation Details

Hyperparameter optimization experiments were conducted using a system equipped with an NVIDIA GeForce RTX 3050 GPU with 4 GB VRAM, a 12th Generation Intel Core i5 processor, and 8 GB RAM. Final model training and evaluation were conducted using an Apple Macbook M1 Generation 1 platform. The framework was implemented using Python 3.10, PyTorch/Torchvision, Ultralytics YOLOv8, and Optuna libraries. The random seed was set at 42 to improve experimental reproducibility. Training was performed for a maximum of 50 epochs with early stopping based on validation loss. All comparative experiments were conducted using identical training, validation, and testing partitions to ensure fair performance evaluation across all models.

III. RESULTS AND DISCUSSION

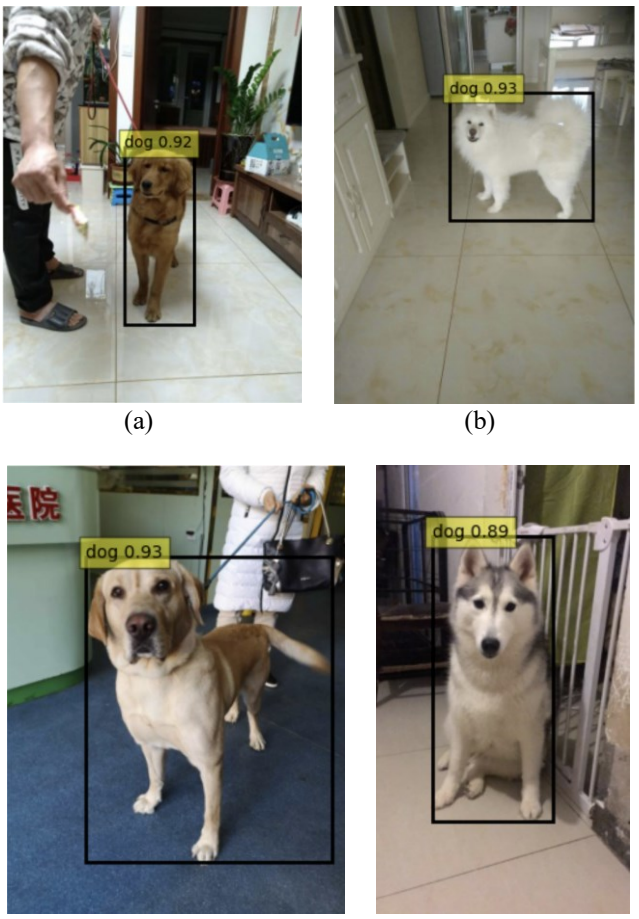
A. YOLOv8n Detection Performance

The object detection performance of the YOLOv8n module was evaluated to assess its effectiveness in localizing dog objects prior to classification. The detection model achieved a precision of 0.9704 and a recall of 0.9504, indicating reliable identification of dog instances with relatively few false positives and missed detections. The model evaluation further produced an mAP@50 of 0.9223 and an estimated mAP@50-95 value of 0.8485, reflecting strong localization capability under the evaluated detection settings.

These results indicate that the YOLOv8n module provides reliable region proposals for downstream classification. The high precision value suggests effective suppression of background regions, while the high recall value indicates that most dog objects were successfully detected prior to classification. This localization stage improves the quality of cropped image regions and reduces background interference before feature extraction by the Vision Transformer classifier.

DOI: 10.34148/teknika.v15i2.1484

The YOLOv8n detection module also demonstrated stable localization performance across multiple dog breed images, as illustrated in Figure 2. Figure 2(a) shows successful localization of a dog object in a Golden Retriever image with a confidence score of 0.92, while Figure 2(b) presents detection results for a Samoyed image with a confidence score of 0.93. Figure 2(c) and Figure 2(d) similarly demonstrate successful localization in Labrador Retriever and Siberian Husky images with confidence scores of 0.93 and 0.89, respectively. These results indicate that the YOLOv8n module reliably localizes dog objects across breeds with varying visual characteristics, including differences in coat texture, facial structure, and body proportions. The localization stage provides reliable object-centered region extraction for the subsection Vision Transformer classification process.



(a) (b) (c) (d)

Figure 2. YOLOv8n Detection Results Demonstrating Successful Dog-Object Localization on Representative Dog Breed Images: (A) Golden Retriever Image With Confidence Score 0.92, (B) Samoyed Image With Confidence Score 0.93, (C) Labrador Retriever Image With Confidence Score 0.93, and Siberian Husky Image With Confidence Score of 0.89.

B. Hyperparameter Optimization

The hyperparameter optimization process identified the best-performing configuration for the Vision Transformer model based on validation accuracy. The optimization was conducted using Optuna with TPE Bayesian optimization, which efficiently explored the defined hyperparameter search space and evaluated multiple candidate configurations during training.

The resulting optimal configuration strikes a balance between convergence stability and generalization performance. Table 2 summarizes the best hyperparameter values selected from all evaluated trials. The obtained learning rate value controls a step size of 4.46×10^{-5} applied to gradient updates during each backpropagation step, enabling gradual parameter refinement and stable convergence of the Vision Transformer parameters. The weight decay value acts as a regularization term by penalizing model weights with a magnitude scaled by 0.0185, thereby reducing excessively large parameter values and mitigating overfitting, which improves generalization.

Table 2. Optimal hyperparameter configuration obtained using Optuna-based Bayesian optimization for the YOLOv8n–ViT-B/16 framework.

Parameter	Best Value
Learning rate	4.46×10^{-5}
Weight decay	0.0185
Batch size	32
Label smoothing	0.1138
Freeze ratio	0.7318
Optimizer	AdamW

The batch size value defines 32 samples processed per training iteration, which stabilizes gradient estimation by averaging updates over multiple samples while maintaining efficient utilization of computational resources. The label smoothing value redistributes the target probability mass by assigning a smoothing factor of 0.1138 away from the ground-truth class and distributing it across non-target classes, reducing overconfidence and improving classification robustness. The freeze ratio indicates that 73.18% of pretrained layers remain frozen, allowing only 26.82% to be updated, enabling dataset-specific adaptation while preserving pretrained representations. The optimizer AdamW applies decoupled weight decay during parameter updates, improving regularization effectiveness and maintaining stable convergence behavior in transformer-based training.

The optimal configuration obtained from the optimization process was subsequently used for the final training stage of the YOLOv8n–ViT-B/16 framework.

C. Training Performance

The training process demonstrated stable convergence over 16 epochs, as illustrated by the training curves in Figures 3 and 4.

The training loss, as shown in Figure 3, decreased progressively from approximately 2.09 to approximately 0.58, indicating effective optimization of model parameters. The validation loss exhibited a consistent downward trend, decreasing from approximately 1.56 in the initial epoch and

DOI: 10.34148/teknika.v15i2.1484

stabilizing within the range of approximately 0.60–0.62 during later epochs, with minor fluctuations after epoch 10.

Training accuracy, as shown in Figure 4, increased rapidly from 29.89% in the first epoch to approximately 98.96% in later epochs, while validation accuracy improved from 65.94% to approximately 98.07%, indicating strong learning capacity and effective generalization performance.

The small gap between the training and validation performance, generally within 0.3% to 1.9%, suggests limited overfitting and stable model generalization capability. Slight oscillations in validation loss after epoch 10 indicate minor sensitivity to optimization dynamics under cosine learning rate scheduling.

The early stopping mechanism terminated training at epoch 16 because validation loss did not improve for five consecutive epochs, confirming model convergence.

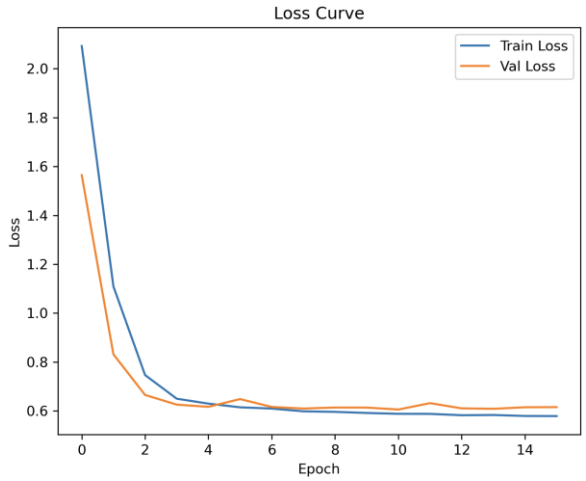


Figure 3. Training And Validation Loss Curves of the Proposed YOLOv8n–Vit-B/16 Framework Over 16 Epochs.

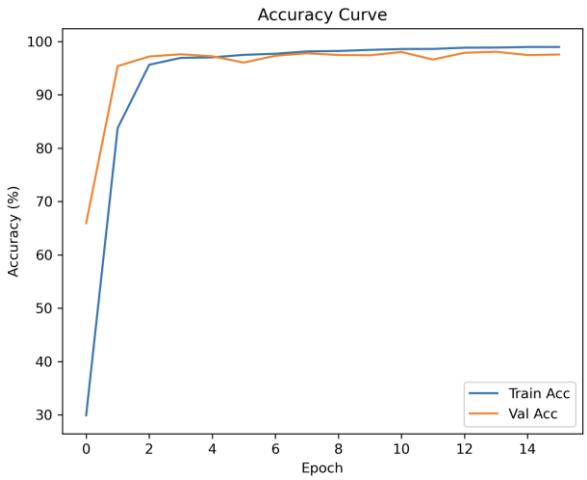


Figure 4. Training and Validation Accuracy Curves of the Proposed YOLOv8n–ViT-B/16 Framework Over 16 Epochs.

D. Confusion Matrix Analysis

The confusion matrix in Figure 5 shows that most predictions are concentrated along the main diagonal, indicating strong classification performance across all dog breed categories. High correct classification rates are observed for most classes, particularly Bernese Mountain Dog, Doberman, German Shepherd, Golden Retriever, and Samoyed, which exhibited near-perfect or highly dominant true positive counts.

A set of misclassification patterns emerges among visually similar breeds, particularly within retriever-related categories. The most notable confusion occurs between Golden Retriever and Labrador Retriever, where 8 Golden Retriever samples are misclassified as Labrador Retriever and 23 Labrador Retriever samples are misclassified as Golden Retriever. This confusion pattern reflects the strong morphological similarity between both retriever breeds, particularly in coat texture, body structure, and facial characteristics. Both classes maintain high precision and recall values despite high inter-class visual similarity conditions.

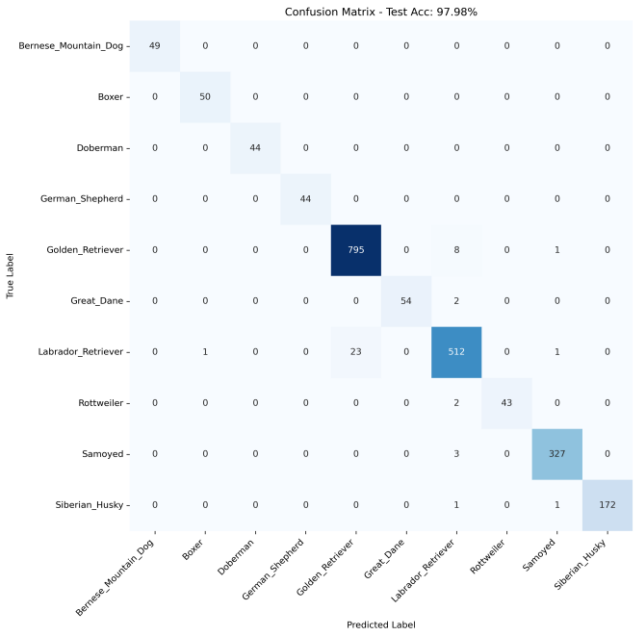


Figure 5. Confusion Matrix of the Proposed YOLOv8n–Vit-B/16 Framework Showing Classification Performance Across Ten Dog Breed Categories Using 2,133 Test Images.

Additional error patterns appeared between Great Dane and Labrador Retriever, where 2 Great Dane samples are classified as Labrador Retriever. Minor confusion is also observed between Rottweiler and Labrador Retriever, with 2 Rottweiler samples incorrectly classified as Labrador Retriever. The Samoyed class exhibited 3 misclassified samples into Labrador Retriever, while the Siberian Husky class showed minimal confusion consisting of 1 sample classified as Samoyed. These patterns indicate that several classification errors are concentrated toward the Labrador Retriever category, which represents one of the largest classes in the dataset and shares overlapping visual characteristics with multiple breeds.

The confusion matrix analysis confirms high classification consistency across all categories, with dominant predictions along the main diagonal. Minor classification patterns persist

DOI: 10.34148/teknika.v15i2.1484

among visually similar breeds, particularly within retriever-type categories, due to overlapping morphological characteristics such as coat texture, facial structure and body proportions. The limited number of off-diagonal errors further indicates stable generalization capability and effective feature representation learning by the proposed YOLOv8n–ViT-B/16 framework.

E. Comparative Performance Evaluation

Comparative experiments were conducted using standalone ViT-B/16 and EfficientNetV2M models to validate the effectiveness of the proposed hybrid framework, as shown in Table 3. All models were trained using identical dataset partitions and utilized identical training, validation, and testing subsets to ensure fair performance comparison. The standalone ViT-B/16 and EfficientNetV2M models received original uncropped images as input, whereas the proposed YOLOv8n–ViT-B/16 framework utilized localized dog regions generated by the YOLOv8n detector prior to classification.

Table 3. Comparative Performance of Different Classification Architectures on the Dog Breed Dataset.

Parameter	Test Accuracy	Macro F1-score	Weighted F1-score
ViT-B/16	0.9709	0.9819	0.9710
EfficientNetV2M	0.9789	0.9816	0.9789
Proposed model	0.9798	0.9876	0.9798

The proposed YOLOv8n–ViT-B/16 framework achieved the highest overall performance across all evaluation metrics. The standalone ViT-B/16 and EfficientNetV2M models also demonstrated strong classification capability across the evaluated dataset. The integration of YOLOv8n-based object localization improved feature focus and reduced background interference prior to classification, contributing to superior overall performance of the proposed framework. Higher macro and weighted F1-score values further indicate improved classification consistency across both majority and minority classes. The comparative results confirm that the object localization stage provides complementary benefits to transformer-based feature representation learning for fine-grained dog-breed classification tasks.

F. Discussion

The quantitative results presented in Table 4 demonstrate the strong overall performance of the proposed YOLOv8n–ViT-B/16 framework across all dog breed categories. The model achieves an overall accuracy of 97.98%, supported by a macro F1-score of 98.76% and a weighted F1-score of 97.98%, indicating balanced performance across both majority and minority classes.

Table 4. Classification Report of the Proposed YOLOv8n–ViT-B/16 Framework on the Dog Breed Test Set

	Prec	Rec	F1	S
Bernese Mt Dog	1.0000	1.0000	1.0000	49
Boxer	0.9804	1.0000	0.9901	50
Doberman	1.0000	1.0000	1.0000	44
German Shepherd	1.0000	1.0000	1.0000	44
Golden Retriever	0.9719	0.9888	0.9803	804
Great Dane	1.0000	0.9643	0.9818	56
Labrador Retriever	0.9697	0.9534	0.9615	537
Rottweiler	1.0000	0.9556	0.9773	45
Samoyed	0.9909	0.9909	0.9909	330
Siberian Husky	1.0000	0.9885	0.9942	174
Accuracy			0.9798	2,133
Macro Avg	0.9913	0.9842	0.9876	2,133
Weighted Avg	0.9799	0.9798	0.9798	2,133

*Prec=Precision, Rec=Recall, F1=F1-score, S=Support

High precision and recall values are consistently observed across most classes, confirming the robustness of the feature representations learned by the Vision Transformer. Bernese Mountain Dog, German Shepherd, and Samoyed classes achieve near-perfect classification scores, with precision and recall values approaching 1.0000. These results indicate that distinctive structural and texture-based features are effectively captured and separated in the learned feature representation space.

Minority classes, including Doberman, German Shepherd, Rottweiler, and Bernese Mountain Dog, maintained strong precision and recall values despite having fewer training samples compared with majority classes such as Golden Retriever and Labrador Retriever. These results suggest that transfer learning, data augmentation, and transformer-based feature representation contributed to stable generalization performance under imbalanced dataset conditions. The consistently high precision and recall values across minority classes indicate limited bias toward majority categories despite the imbalanced class distribution.

Strong performance is also observed in high-support classes, such as Golden Retriever and Labrador Retriever, which achieved F1-scores of 0.9803 and 0.9615, respectively. Minor performance reductions in these classes are associated with inter-class visual similarity, particularly in coat color and facial structure, leading to limited overlap in feature distributions.

Minor classification variation remains observable in the Labrador Retriever, Great Dane, and Rottweiler classes. These cases correspond to the confusion patterns presented in Figure 5 and indicate that fine-grained discrimination remains challenging when inter-class morphological differences are limited, particularly among large-breed and retriever-type categories.

The confusion matrix in Figure 5 complements the statistical results in Table 4 by illustrating the distribution of correct and incorrect predictions across all classes. Misclassifications are primarily concentrated among breeds with high visual similarity, while structurally distinct breeds maintain near-perfect classification consistency.

The YOLOv8n detection module contributes significantly to classification stability by isolating dog-specific regions prior to feature extraction. This reduces background interference and ensures that the Vision Transformer processes only

DOI: 10.34148/teknika.v15i2.1484

semantically relevant regions. The ViT-B/16 architecture further enhances discriminative capability by modeling long-range dependencies through self-attention, enabling effective aggregation of global contextual information.

The self-attention mechanism enables the Vision Transformer to model long-range spatial relationships among multiple regions of the dog image, including facial structure, coat texture, ear shape, and body proportions. This capability is particularly beneficial for fine-grained breed classification, where discriminative visual characteristics are distributed across different spatial locations rather than concentrated in a single local region. Global contextual representation learning allows the model to capture relationships between distant image patches, which supports discrimination among visually similar breeds such as Golden Retriever and Labrador Retriever.

The combined analysis of Table 4 and Figure 5 confirms that classification errors are not caused by training instability or feature collapse. The remaining errors are primarily attributable to intrinsic inter-class similarity among certain dog breeds rather than limitations in the model's learning capacity.

The integration of YOLOv8n and Vision Transformer demonstrates complementary strengths, with object detection improving input quality and transformer-based encoding enhancing semantic discrimination. This synergy results in strong generalization performance across diverse dog-breed categories while maintaining robustness to class imbalance and visual-similarity challenges.

The comparative evaluation against standalone ViT-B/16 and EfficientNetV2M further demonstrates the effectiveness of integrating YOLOv8n localization with transformer-based classification. All evaluated models achieved high classification accuracy, but the proposed framework consistently obtained superior macro and weighted F1-scores. These improvements indicate that object localization contributes to more discriminative feature extraction by reducing irrelevant background information prior to transformer-based representation learning. The higher macro F1-score achieved by the proposed framework further indicates improved classification consistency across minority classes and visually similar breed categories.

IV. CONCLUSION

This study proposed a Vision Transformer-based framework for dog breed classification using the ViT-B/16 architecture integrated with YOLOv8n object detection. The proposed YOLOv8n–ViT-B/16 framework achieved an accuracy of 97.98% on 2,133 testing images, indicating strong capability in distinguishing visually similar dog breeds. Comparative experiments against standalone ViT-B/16 and EfficientNetV2M further demonstrated that the integration of YOLOv8n localization improves classification consistency and feature discrimination in dog breed recognition tasks.

Transformer-based visual representation learning enabled the model to capture global contextual relationships and subtle discriminative features effectively. The integration of YOLOv8n for object localization ensured that only dog-relevant regions were processed, reducing background interference and improving feature quality. Transfer learning and hyperparameter optimization further enhanced classification robustness and generalization performance.

The combined YOLOv8n–ViT-B/16 architecture provides a complementary learning mechanism, in which object detection improves the quality of input regions, and the Vision Transformer performs high-level semantic modeling. This design contributes to stable performance across diverse breed categories, including visually similar classes. Comparative evaluation results further indicate that the proposed hybrid framework provides more consistent classification performance than standalone transformer-based and convolutional architectures on the evaluated dog-breed dataset.

Several limitations remain in this proposed framework, including class imbalance, high inter-class visual similarity among certain breeds, and dependence on variations in image quality. Future work may explore lightweight transformer architectures, hybrid CNN-transformer models, explainable artificial intelligence techniques, and larger-scale datasets to further improve classification efficiency, robustness, and real-time deployment capability.

REFERENCES

[1] N. Atero *et al.*, "An assessment of the owned canine and feline demographics in Chile: registration, sterilization, and unsupervised roaming indicators," *Prev. Vet. Med.*, vol. 226, p. 106185, May 2024, doi: 10.1016/j.prevetmed.2024.106185.

[2] H. T. Yim, K. J. Flay, O. Nekouei, P. V. Steagall, and J. A. Beatty, "Pet Dog Choice in Hong Kong and Mainland China: Exploring Owners' Motivations, Behaviours, and Perceptions," *Animals*, vol. 15, no. 4, p. 486, Feb. 2025, doi: 10.3390/ani15040486.

[3] Z. Malinovská and E. Čonková, "Genes of Congenital Dermatologic Disorders in Dogs—A Review," *Folia Vet.*, vol. 65, no. 4, pp. 38–46, Dec. 2021, doi: 10.2478/fv-2021-0036.

[4] W. D. Mansilla, L. Fortener, J. R. Templeman, and A. K. Shoveller, "Adult dogs of different breed sizes have similar threonine requirements as determined by the indicator amino acid oxidation technique," *J. Anim. Sci.*, vol. 98, no. 3, Mar. 2020, doi: 10.1093/jas/skaa066.

[5] M. R. Viant, C. Ludwig, S. Rhodes, U. L. Günther, and D. Allaway, "Validation of a urine metabolome fingerprint in dog for phenotypic classification," *Metabolomics*, vol. 5, no. 4, pp. 517–517, Dec. 2009, doi: 10.1007/s11306-009-0172-4.

[6] G. Perez, Y. He, Z. Lyu, Y. Chen, N. R. Howe, and H. M. Rando, "Standardizing canine breed data in veterinary records is challenging, but computer vision offers an alternative perspective on breed assignment," *Am. J. Vet. Res.*, vol. 86, no. S1, pp. S38–S45, 2025, doi: 10.2460/ajvr.24.10.0315.

[7] C. L. Cazer, P. Basran, and R. Ivanek-Miojevic, "From bark to bytes: artificial intelligence transforming veterinary medicine," *Am. J. Vet. Res.*, vol. 86, no. S1, pp. S4–S5, 2025, doi: 10.2460/ajvr.86.s1.editorial.

[8] A. Dabrowski, K. Lichy, P. Lipiński, and B. Morawska, "Dog Breed Library with Picture-Based Search Using Neural Networks," in *2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT)*, 2021, pp. 17–20. doi: 10.1109/CSIT52700.2021.9648628.

[9] A. Nawaz, R. S. Shoukat, M. Shehab, K. El Hindi, and Z. Ahmed, "CLIP-ASN: A Multi-Model Deep Learning Approach to Recognize Dog Breeds," *Computers, Materials & Continua*, vol. 85, no. 3, pp. 4777–4793, 2025, doi: 10.32604/cmc.2025.064088.

[10] P. O. Adejumobi, I. O. Adejumobi, O. A. Adebisi, S. O. Ayanlade, and I. I. Adeaga, "Automatic classification of breeds of dog using convolutional neural network," *Nigerian Journal of Technological Development*, vol. 20, no. 3, pp. 199–209, Oct. 2023, doi: 10.4314/njtd.v20i3.1485.

[11] Y. A. Reddy, Y. S. Kumar, S. M, and S. C. Mana, "Dog Breed Identification using ResNet Model," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 7s, pp. 64–71, Jul. 2023, doi: 10.17762/ijritcc.v11i7s.6977.

[12] P. Borwarnginn, K. Thongkanchorn, S. Kanchanapreechakorn, and W. Kusakunniran, "Breakthrough Conventional Based Approach for Dog Breed Classification Using CNN with Transfer Learning," in *2019 11th International Conference on Information Technology and Electrical Engineering (ICITEE)*, 2019, pp. 1–5. doi: 10.1109/ICITEED.2019.8929955.

[13] T. P. G. James, P. C, S. S, G. Malathi, K. S, and N. Venu, "Efficient Canine Vision: Accurate Dog Breed Classification with EfficientNet," in *2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies*, 2024, pp. 1–6. doi: 10.1109/TQCEBT59414.2024.10545139.

[14] B. Palanisamy *et al.*, "Transformers for Vision: A Survey on Innovative Methods for Computer Vision," *IEEE Access*, vol. 13, pp. 95496–95523, 2025, doi: 10.1109/ACCESS.2025.3571735.

[15] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[16] C. Tang, Y. Zhou, W. Qin, Y. Zhang, R. Wu, and W. Wang, "Transformer-Driven Self-Supervised Learning for Visual Understanding: Methods and Applications," in *2025 4th International Conference on Artificial Intelligence, Internet of Things and Cloud Computing Technology (AIoTC)*, IEEE, Aug. 2025, pp. 155–159. doi: 10.1109/AIoTC66747.2025.11198691.

[17] M. Filipiuk and V. Singh, "Comparing Vision Transformers and Convolutional Nets for Safety Critical Systems," in *SafeAI@AAAI*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247321083

[18] J. Maurício, I. Domingues, and J. Bernardino, "Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review," *Applied Sciences*, vol. 13, no. 9, p. 5521, Apr. 2023, doi: 10.3390/app13095521.

[19] V. H. B. Canto, J. R. R. Manesco, G. B. de Souza, and A. N. Marana, "Dog Face Recognition Using Vision Transformer," in *Intelligent Systems*, M. C. Naldi and R. A. C. Bianchi, Eds., Cham: Springer Nature Switzerland, 2023, pp. 33–47.

[20] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[21] D.-N. Zou, S.-H. Zhang, T.-J. Mu, and M. Zhang, "A new dataset of dog breed images and a benchmark for finegrained classification," *Comput. Vis. Media (Beijing).*, vol. 6, no. 4, pp. 477–487, Dec. 2020, doi: 10.1007/s41095-020-0184-6.

[22] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jul. 2017, pp. 936–944. doi: 10.1109/CVPR.2017.106.

[23] Z. Liu, P. Gong, and J. Wang, "Attention-Based Feature Pyramid Network for Object Detection," in *Proceedings of the 2019 8th International Conference on Computing and Pattern Recognition*, New York, NY, USA: ACM, Oct. 2019, pp. 117–121. doi: 10.1145/3373509.3373529.

[24] X. Liu, H. Pan, and X. Li, "Object detection for rotated and densely arranged objects in aerial images using path aggregated feature pyramid networks," in *MIPPR 2019: Pattern Recognition and Computer Vision*, Z. Liu, J. K. Udupa, N. Sang, and Y. Wang, Eds., SPIE, Feb. 2020, p. 27. doi: 10.1117/12.2538090.

[25] A. Vaswani *et al.*, "Attention Is All You Need," Aug. 2023.

[26] H. M. Khan, A. Khan, S. G. Villar, L. A. D. Lopez, A. Almaleh, and A. M. Al-Qahtani, "A Comparative Study of Optimized-LSTM Models Using Tree-Structured Parzen Estimator for Traffic Flow Forecasting in Intelligent Transportation," *Computers, Materials & Continua*, vol. 83, no. 2, pp. 3369–3388, 2025, doi: 10.32604/cmc.2025.060474.

[27] T. Vaiyapuri, "An Optuna-Based Metaheuristic Optimization Framework for Biomedical Image Analysis," *Engineering, Technology & Applied Science Research*, vol. 15, no. 4, pp. 24382–24389, Aug. 2025, doi: 10.48084/etasr.11234.