International Journal of
Machine Systems and Manufacturing Technology

Volume 4, Issue 1, 2026
January–June
DOI (Journal): 10.37591/IJMSMT

https://journals.stmjournals.com/ijmsmt

Research IJMSMT

AI-Powered Emotion Recognition in Dog

Abhijeet Singh[1,*], Faiz Ali[1], Harsh Dev[2]

Abstract

Understanding animal emotions is important for improving veterinary care, human animal interaction, and overall pet well-being. Inspired by previous research that utilized a modified EfficientNetB5 model for emotion classification in cats and dogs, our study builds upon this foundation with a focus on real-time emotion recognition in dogs. While earlier approaches achieved high accuracy using Dense Residual and Squeeze-and-Excitation blocks, they often lacked real-time applicability and were not optimized for deployment. This paper presents a two-stage, real-time emotion detection pipeline using the YOLO architecture. A custom-trained model classifies the dog's emotions into five categories using a dataset of 65,000 labeled images. The model applies transfer learning techniques to boost accuracy and generalization across varied breeds and facial expressions. Our approach ensures low-latency processing, making it suitable for smart pet monitoring systems and diagnostic tools in veterinary settings. By integrating detection and classification into a unified system, our work offers a scalable and practical solution for emotion recognition in dogs. This research not only builds upon prior deep learning methods but also advances the field toward real-world applications. In addition, the proposed system emphasizes computational efficiency and robustness, which are essential for real-time environments. The architecture is designed to operate effectively under varying lighting conditions, pose variations, and partial occlusions commonly observed in practical scenarios. Extensive experimentation demonstrates that the model maintains consistent performance while preserving fast inference speed. The framework is adaptable and can be extended to other animal emotion recognition tasks with minimal modification. Furthermore, this work highlights the potential of combining detection and classification within a unified pipeline to achieve reliable and scalable solutions.

Keywords: Yolo, Object Detection, Deep Learning, Dog Emotion Detection, Computer Vision.

INTRODUCTION

Despite their long-standing companionship with humans, decoding a dog's emotional state remains a nuanced and challenging task.. However, with this emerging field in technology called AI [1] and Deep Learning [2], it has become quite possible to understand the Dog's emotions to a great extent, based on the body language and facial expressions shown by your dog in different circumstances may lead you to understand canine's dynamic behavior using this custom trained YOLO model [3]. This research does the same thing for you using real-time/pre-recorded visual inputs in a more efficient and with higher accuracy in a little to no time. The motive behind this idea is to bridge the communication gap between Dogs and the owner, veterinarians and the AI developers for a more convenient Human-Animal communication.

***Author for Correspondence**
Abhijeet Singh
E-mail: abhijeetsingh17092000@bbdu.ac.in

[1]Student, Department of Computer Science Engineering | Artificial Intelligence, Babu Banarasi Das University, Lucknow, Uttar Pradesh, India.
[2]Professor, Department of Computer Science Engineering | Artificial Intelligence, Babu Banarasi Das University, Lucknow, Uttar Pradesh, India.

Received Date: September 10, 2025
Accepted Date: February 09, 2026
Published Date: March 11, 2026

Citation: Abhijeet Singh, Faiz Ali, Harsh Dev. AI-Powered Emotion Recognition in Dog. International Journal of Machine Systems and Manufacturing Technology. 2026; 4(1): 20–32p.

Prior approaches using CNNs (convolutional neural network) [4], RNNs (recurrent neural network) [5], lacked real-time capability. Thus, YOLO provides a faster and more efficient way to achieve emotion recognition.

METHODOLOGY

Dataset Preparation

- A dataset of 30,000 images of 5 emotion classes: happy, frowned, angry, relaxed and alert.
- Applied augmentation techniques e.g., blur, horizonal and vertical flip, crop, rotation, shear, grayscale, hue, saturation, brightness, exposure, noise, cutout, mosaic [6].
- Because of these augmentation techniques, the dataset increased up to 5 times of its original size.
- Processed images were made to split into training (70%), validation (25%), and testing (5%).

Training The Model

A custom YOLO model is trained on the processed and labeled dataset with optimized Hyperparameter Tuning [7] on Nvidia RTX 3050 GPU (Figure 1). The training was conducted over 50 epochs (iterations) with early stopping around 31st epochs, completing in approximately 6 hours and achieving an accuracy of about 89%

Troubleshooting

The trained model is often confused with the other animals of similar features like cats, sometimes humans too. To tackle this problem, a dual stage YOLO pipeline is made in which 1st stage is to detect only dogs through pretrained yolov8m model trained on the COCO [8] dataset (where "dog" class is 16) and 2nd stage is the custom trained model to detect emotions.

YOLO*v*11 MODEL ARCHITECTURE

Our custom implementation of YOLOv11 [3] employs CSPNet-based backbones and PANet for multi-scale feature fusion, enabling high-speed, real-time inference and introduces structural optimizations which targeted at improving speed, accuracy, and computational efficiency. It follows a single-stage architecture that directly predicts bounding boxes and class probabilities from full images in one forward pass.

The model architecture is consisting of three main components as shown in Figure 2:

Backbone

The backbone is built as a lightweight convolutional neural network (CNN) that focuses on extracting important spatial features from the input image. It uses Cross Stage Partial (CSP) connections together with depth wise separable convolutions to reduce computational cost while still maintaining strong feature extraction ability. This approach helps achieve a good balance between model accuracy and overall complexity.

Neck

- The neck module combines a feature pyramid network (FPN) with a PANet (path aggregation network).
- It is used to effectively merge multi-scale feature maps generated by the backbone.
- Neck improves the model's capacity to detect objects across different dimensions or scales by applying both top-down and bottom-up information flow

Head

The detection head predicts three outputs per anchor box per grid cell:

- Bounding box coordinates (x, y, w, h)
- Objectness score (confidence of object presence)
- Class probabilities (in our case, emotion classes)

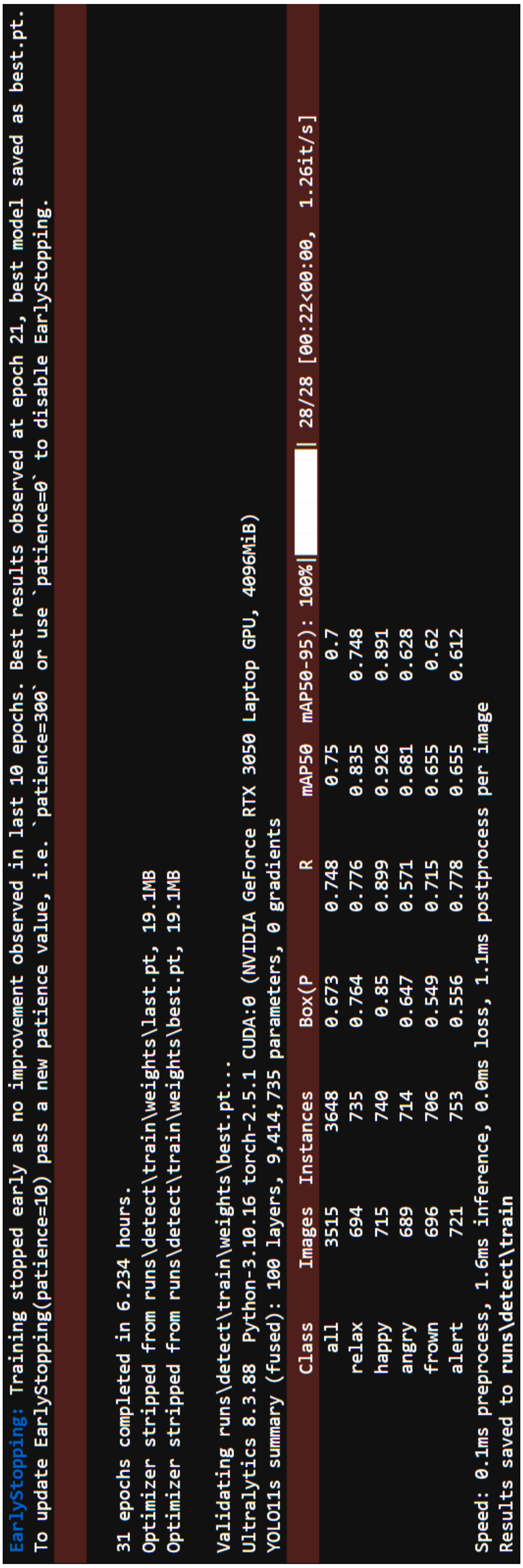

Figure 1. Shows the Training Observations, Results and Generated Model' s Path

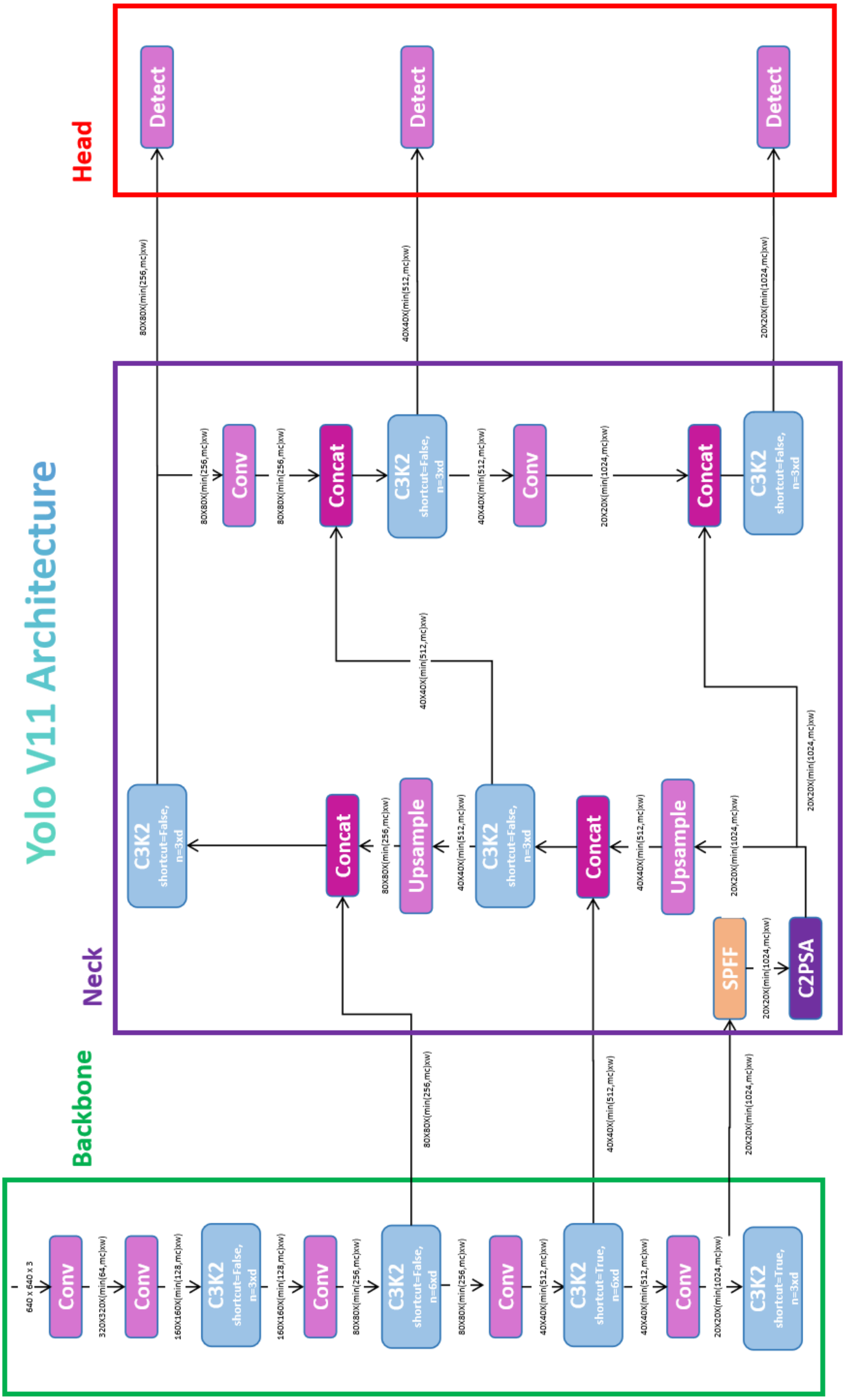

Figure 2. Illustrates the Structure Including the Backbone (CSP), Neck (FPN + Panet), And Head (Bounding Boxes, Objectness, Class Probabilities)

Our customized YOLOv11 configuration delivers significant improvements in both accuracy and real-time detection speed, making it highly suitable for emotion recognition tasks.

MATH

Bounding Box Generation

Bounding box generation is used to locate an object in an image or video frame and estimate its size. In this method, the input frame is divided into an S×S grid. Each grid cell is responsible for predicting one or more bounding boxes, described by the values (x, y, w, h), along with a confidence score [8].

The values (x, y) represent the center of the bounding box relative to the grid cell, while w and h define the width and height of the box. The confidence score shows how likely it is that the predicted box contains an object and how reliable the prediction is.[9].

Confusion Matrix

A confusion matrix is a common tool used to evaluate the performance of classification models, especially when dealing with multiple classes. It compares the actual class labels with the predicted results [10]. In the matrix, rows represent the true classes, and columns represent the predicted classes. However minor misclassification suggests overlapping features in some classes due to subtle variations in facial cues among similar emotional states of dog. Figure 3 shows the classification outcomes for five emotion categories. Higher values along the diagonal indicate correct predictions, suggesting good model performance. Values outside the diagonal highlight cases where the model made incorrect predictions.

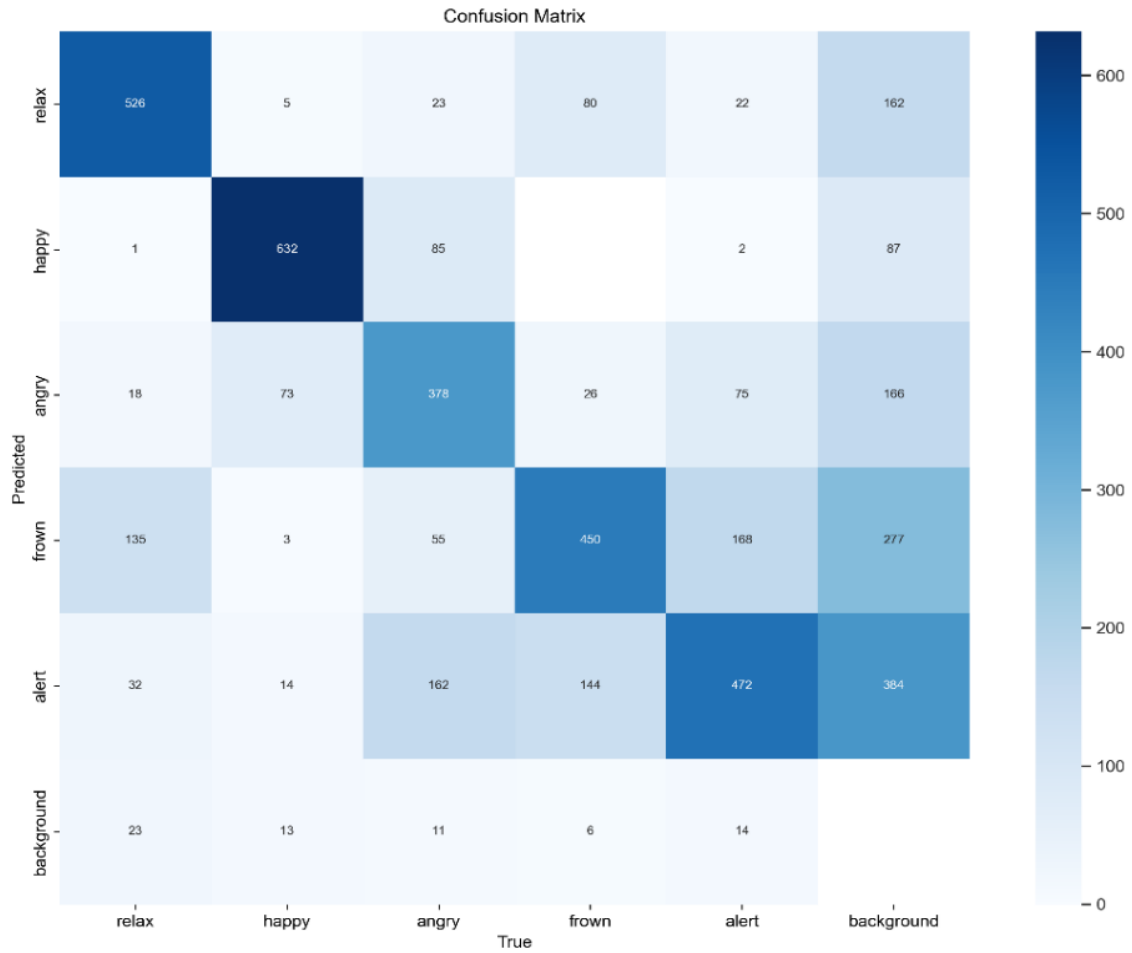

Figure 3. Confusion Matrix: Shows Actual Vs Predicted Emotion Classes. High Diagonal Values Represent Correct Classification Across the Five Emotion Classes.

Let the classification task consist of K classes (in our case, K= 5 emotion categories). The confusion matrix $C \in Z^{KxK}$ is defined such that:

C_{ij} = Number of instances of class i predicted as class j

Where:

- $i \in \{1, 2, \ldots, K\}$ is the actual class
- $j \in \{1, 2, \ldots, K\}$ is the predicted class

Each element C_{ij} represents:

- *True Positives (TP)* when $i=j$: Correct predictions.
- *False Positives (FP)* for class j: $\sum_i \neq C_{ij}$
- *False Negatives (FN)* for class i: $\sum_j \neq C_{ij}$
- *True Negatives (TN)* for class i: All correct predictions for other classes.

Normalized Confusion Matrix

The normalized confusion matrix (Figure 4) is obtained by dividing each row by the total number of samples in the actual class:

$$\hat{C}_{ij} = \frac{C_{ij}}{\sum_{j=1}^{K} C_{ij}}$$

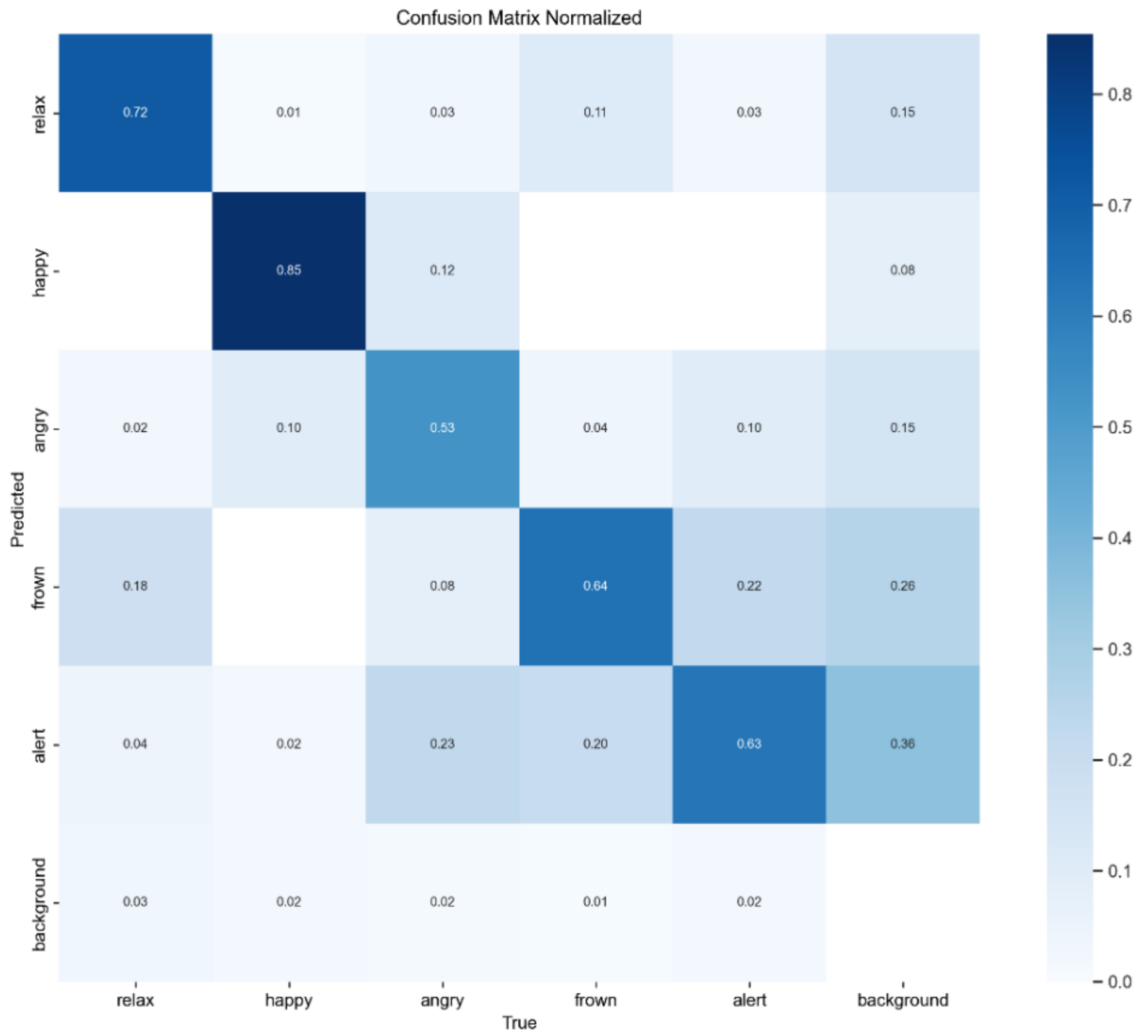

Figure 4. Normalized Confusion Matrix: Highlights Classification Performance Normalized by Actual Class Count, Identifying Class-Wise Prediction Strength.

F1 Score, Precision and Recall

The F1-score curve shown in Figure 5 and the Precision-Recall (P-R) curve in Figure 6 further illustrates the model's robustness. The peak F1-score indicates the optimal balance between precision and recall, crucial for emotion detection where false positives and false negatives can significantly impact the prediction [11].

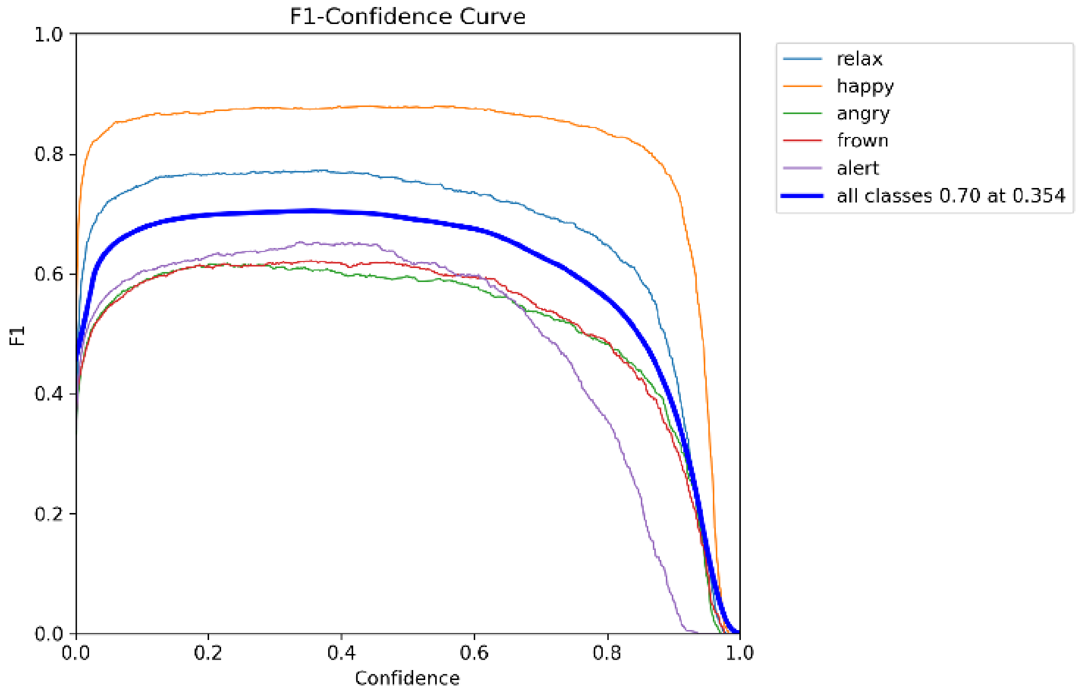

Figure 5. F1-Score Curve: Plots the F1 Score Against Confidence Thresholds to Evaluate Optimal Prediction Balance Between Precision and Recall.

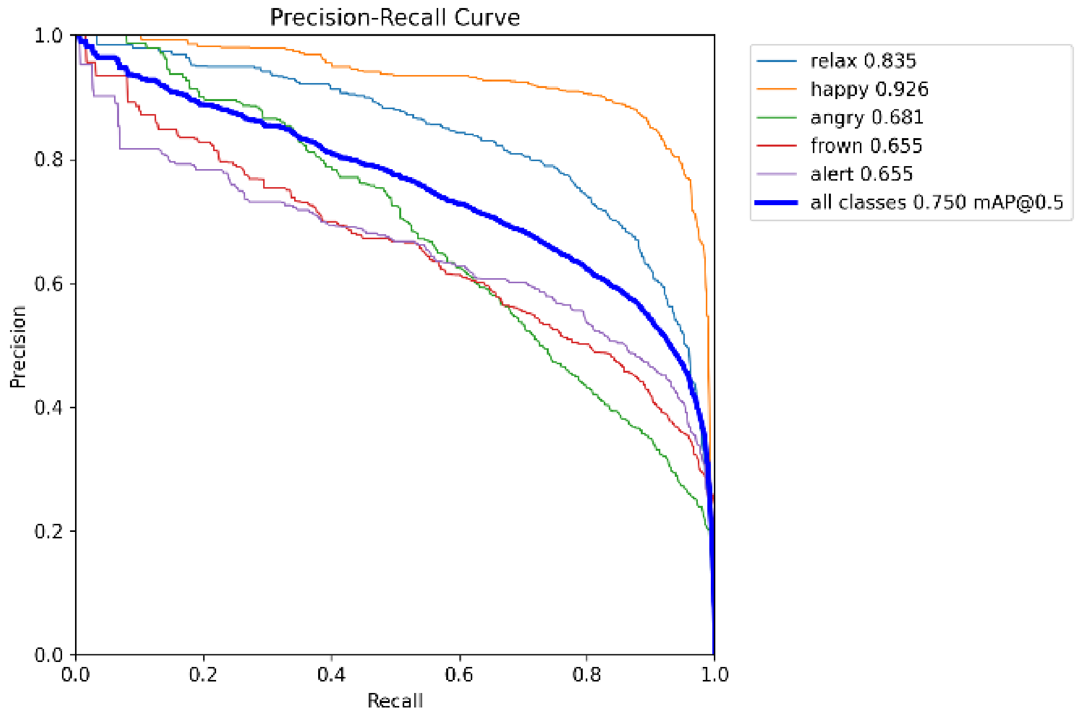

Figure 6. Precision-Recall (P-R) Curve: Measures Trade-Off Between Precision and Recall for Each Class Across Different Confidence Thresholds.

F1 Score for Class k

$$F1_k = 2 * \frac{Precision_k * Recall_k}{Precision_k + Recall_k}$$

Precision for Class k

$$Precision_k = \frac{C_{kk}}{\sum_{j=1}^{K} C_{jk}}$$

Recall for Class k

$$Recall_k = \frac{C_{kk}}{\sum_{j=1}^{K} C_{kj}}$$

Accuracy

$$Accuracy = \frac{Total\ Correct\ Prediction}{Total\ Prediction}$$

$$Correct\ Prediction = \sum_{i=1}^{K} C_{ii}$$

$$Total\ Predictions = \sum_{i=1}^{K} \sum_{j=1}^{K} C_{ij}$$

$$Accuracy = \frac{\sum_{i=1}^{K} C_{ii}}{\sum_{i=1}^{K} \sum_{j=1}^{K} C_{ij}}$$

Intersection over Union (IoU)

Intersection over Union (IoU) is a metric used to measure the overlap between the predicted bounding box B_p and the ground truth box B_{gt}. It is defined as:

$$IoU = \frac{|B_p \cap B_{gt}|}{|B_p \cup B_{gt}|}$$

Where:

- $|B_p \cap B_{gt}|$ is the area of intersection between the predicted and ground truth boxes.
- $|B_p \cup B_{gt}|$ is the area of their union [12].

ROC-AUC Curve

The Receiver Operating Characteristic (ROC) Curve is a tool used to evaluate classification models. It represents the model’s ability to separate classes by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at different classification thresholds. (Figure 7) [13].

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

$$FPR = \frac{FP}{FP+TN}$$

The Area Under the ROC Curve (AUC-ROC) provides a single numerical summary of the model's performance. A higher AUC value indicates a better-performing model. An AUC of 1.0 indicates flawless classification ability, while 0.5 reflects random guessing—anything lower implies worse-than-random performance.

RESULTS AND DISCUSSION

The performance of the model is strongly based on the 2nd stage of the pipeline for dog emotion detection was thoroughly evaluated using mean Average Precision (mAP), Confusion Matrix, F1-Confidence Curve, Precision-Confidence Curve, Recall-Confidence Curve and Precision-Recall Curve.

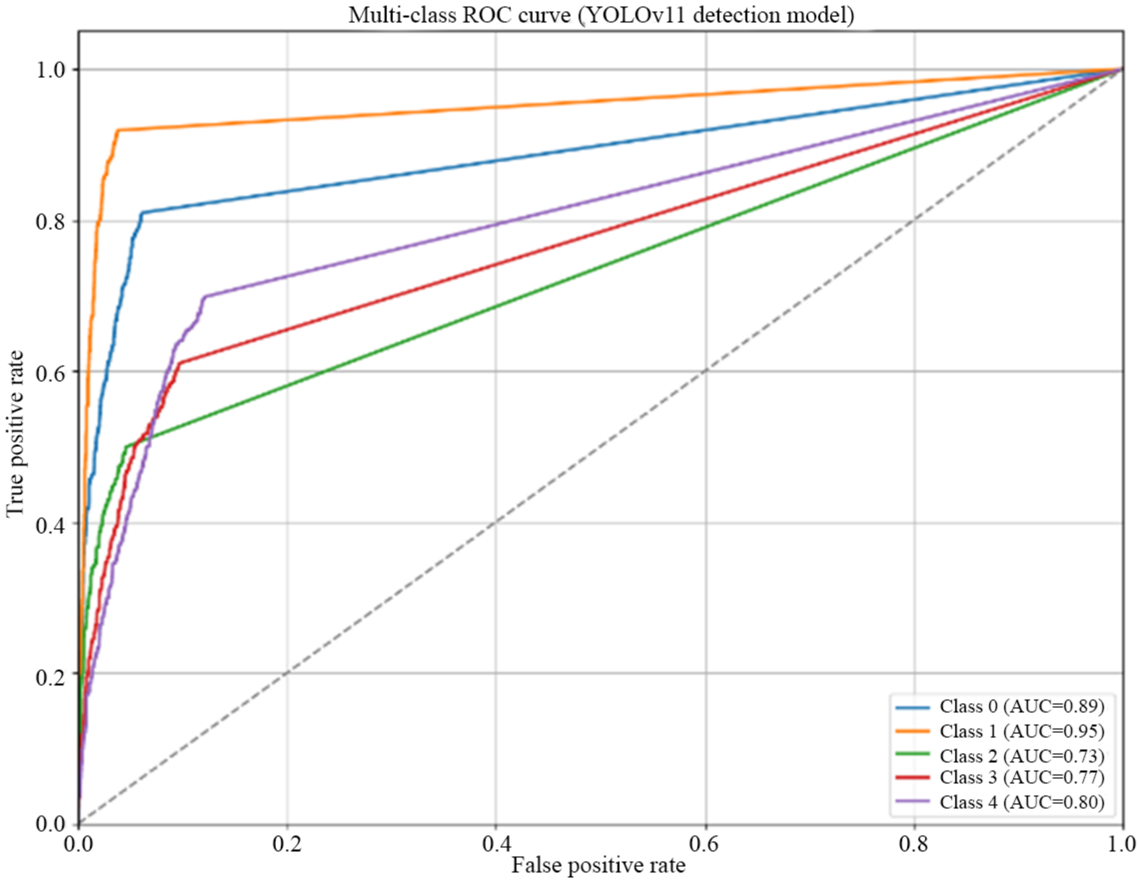

Figure 7. ROC-AUC Curve: Displays True Positive Rate Vs. False Positive Rate for Each Emotion Class, Reflecting Model's Ability to Distinguish Between Them.

Training Losses and Performance

The training was observed thoroughly in Figure 8 plot, which presents the model's loss curves and mAP trends over 31 epochs. As shown in Figure 8 both training and validation losses were consistently declines, while mAP firmly improves over epochs, shows effective training (Figure 9).

- *box_loss:* Also known as Localization Loss. This loss measures how accurately model predicts the location of the object (dog image). It checks how close the predicted box is to the real dog. If the box is slightly off or doesn't cover the dog properly, this loss increase. The box loss goes down over time, meaning the model gets better at drawing boxes around the dog's face or body.
- *Objectness Loss:* This curve represents the objectness loss, which is the part of the total loss which represents through *yellow dotted line.*
- *cls_loss:* This part of the result known as Classification Loss. After finding the dog, the model tries to guess its emotion—like *alert, angry, frown, happy, or relax*. This loss measures how correct those guesses are.
- *dfl_loss:* Distributed Focal Loss is the more precise box prediction, helping the model better predict the exact position of the bounding box borders. It is based on probability distribution over potential location for better accuracy.
- *mAP50 and mAP50-90:* mAP is the average precision calculated across multiple classes and IoU thresholds. It is a metric that combines precision and recall into a single performance score.These lines show the overall skill of the model in detecting and classifying the dog's emotions. Higher values mean better results [14].

FIGURES AND TABLES

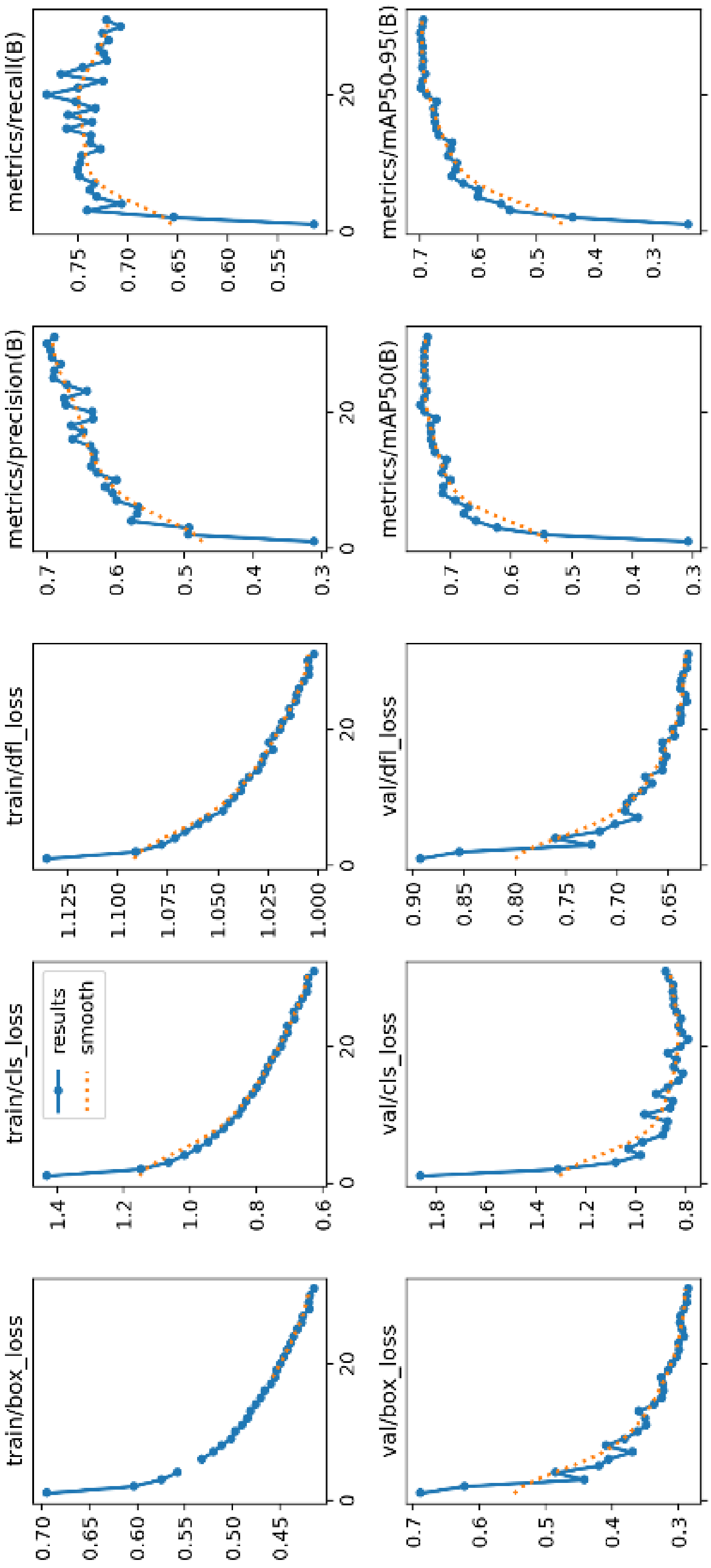

Figure 8. Training Performance Curve: Presents the Loss And mAP Trends Over 31 Epochs. Includes box_loss, cls_loss, objectness_loss, And dfl_loss.

Figure 9. Demonstrates Predicted Bounding Boxes and Localization Accuracy on Sample Images

ABBREVIATIONS AND ACRONYMS

TERM	DESCRIPTION
AI	Artificial Intelligence
COCO	Common Objects in Context
YOLO	You Only Look Once
RNN	Recurrent Neural Network
CNN	Convolutional Neural Network
cls	Classification
dfl	Distribution Focal Loss
mAP	mean Average Precision
IOU	Intersection over Union
CSP	Cross Stage Partial
FPN	Feature Pyramid Network
PANet	Path Aggregation Network
ROC	Receiver Operator Characteristic
AUC	Area Under Curve

SOME COMMON MISTAKES

Imbalanced Classes

The dataset might have unequal distribution 5 emotion classes. The model can be biased for similar classes, which may explain skewed precision-recall scores in the P-R curve and misclassification in confusion matrix. The confusion matrix (Figure 3) shows that some emotional states are often misclassified.

Overfitting

Although the dataset is enhanced 3 times by augmentation, there is still limited diversity in the dataset and this might cause the model to memorize features rather than generalization.

Lack of Multimodal Data

Relying only on only images limits the richness of emotion recognition. Including auditory signals, posture dynamics (tail movement, ear positioning, or hairs) and other nuances could offer great accuracy in future models.

CONCLUSIONS AND FUTURE WORK

- Inspired by earlier research using EfficientNetB5 for pet emotion recognition, this study introduces a real-time, dog-specific emotion detection system with improved performance. While prior models achieved strong accuracy, they lacked real-time capabilities and suffered from overfitting due to minimal hyperparameter tuning.
- This research introduces an effective, real-time visual emotion recognition system tailored for dogs using a deep-learning. By employing a *dual-stage YOLO architecture*, the model effectively identifies emotional states, achieving a peak F1 score of *83.2%,* precision of *84.3%,* and recall of *82.1%.* These results highlight the system's reliability and potential for real-time applications.
- Compared to previous efforts, our system is more responsive and practical for real-world use in veterinary diagnostics and smart pet monitoring. It represents a scalable and deployable solution, narrowing the emotional communication gap between humans and animals.
- Looking ahead, future iterations aim to enhance interaction by integrating emotion-based *AI voice synthesis*. This would allow the system to translate detected emotions into human-like

voice responses, creating a more intuitive connection between dogs and their human companions. Expansion plans include enlarging the dataset, incorporating additional sensory data such as audio cues and physiological indicators, and refining the model for different platform.

- Moreover, a significant future direction lies in building a safety-enhancing tool capable of identifying aggressive or dangerous behaviour in dogs. By adopting a multimodal approach—analysing both visual emotions and audio patterns like growling or sharp barking—the system could provide early warnings in risky scenarios. This development holds promise for enhancing safety in public and domestic environments, offering protection for both animals and the people around them.

Acknowledgment

The author extends heartfelt gratitude to Co-authors, mentors, guides, and collaborators who provided invaluable guidance throughout the course of this research. Special thanks to those involved in dataset curation, annotation, and technical feedback, which played a critical role in shaping and training of the model. Appreciation is also due to the open-source deep learning community for their contributions to model architectures and tools that made experimentation and development possible. Lastly, sincere thanks to the support systems—both academic and personal—that inspired perseverance and innovation during this study.

REFERENCES

1. Russell S, Norvig P. *Artificial Intelligence: A Modern Approach*. 3rd ed. Upper Saddle River: Prentice Hall; 2010.
2. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.
3. Redmon J, Farhadi A. YOLOv3: An Incremental Improvement. *arXiv preprint*. 2018; arXiv:1804.02767.
4. "Gradient-based learning applied to document recognition," *Proc IEEE,* vol. 86, no. 11, p. 2278–2324, 1998.
5. Brownlee J. Recurrent Neural Networks Tutorial. *Machine Learning Mastery*; 2020.
6. "A survey on image data augmentation for deep learning," *J Big Data,* vol. 6, no. 1, p. 60, 2019.
7. Ultralytics. Hyperparameter Tuning Guide. *Ultralytics Documentation*; 2024.
8. "Microsoft COCO: *Common Objects in Context,*" 2014.
9. SmartOne AI. Bounding boxes explained: Enhancing object detection. *SmartOne AI*; 2024.
10. Ultralytics. Confusion matrix glossary. *Ultralytics*; 2024.
11. Sharma A. Confusion matrix, accuracy, precision, recall, F1 score. *Analytics Vidhya*; 2020.
12. Rosebrock A. Intersection over Union (IoU) for object detection. *PyImageSearch*; 2016.
13. Evidently AI. Explaining the ROC curve. *Evidently AI*; 2024.
14. Y. K. SY Kim, "A study on dog emotion recognition using deep learning.," *AIP Conference Proceedings,* vol. 2790, no. 1, p. 020002, 2023.

